Molecular evolution of nuclease domains in nucleotide polymerases

(核酸ポリメラーゼに存在するヌクレアーゼドメインの分子進化)

Tsuyoshi Shirai

白井剛

主論文

Molecular evolution of nuclease domains in nucleotide polymerases

Tsuyoshi Shirai

Contents

Abstract

Nucleotide polymerases are large and complex proteins by reflecting their essential and complicated activities in organisms such as gene transcription or genome replication. Proteins of complex functions and structures are thought to have evolved by combining smaller functional/structural units, such as modules or domains. The molecular mechanism of structural reorganization in protein is one of the most important subjects in molecular evolution.

In this study, a part of DNA-directed RNA polymerase is shown to be homologous with microbial ribonuclease (RNase: *e.g.* barnase from *Bacillus amyloliquefaciens*) and named as RNase-like domain. From a comparison of amino acid sequences of the RNases and the RNA polymerases on three-dimensional structure of barnase, active sites and key residues in structure formation (residues at hydrophobic cores and turns) of the RNases are shown to be conserved in the RNase-like domains. Compatibility of the domain sequences to the RNase structure was tested on molecular models of the domain which were constructed by a method of homology-modeling. The results show that the most part of the models, except for few regions, are stable and have proper solvent accessibility for a globular domain.

The finding of RNase-like domain suggests that the RNA polymerase has acquired an RNase by domain fusion in the course of evolution. The domain might work in proof reading in mRNA transcription as the nuclease domain of DNA polymerase does in replication. Binding of nascent RNA is also a putative function of the domain. In fact, the RNA cleavage activity of the RNA polymerases was reported by the other researchers soon after the finding of the domain. The RNase-like domain is a candidate of the catalytic center of the cleavage activity. This is the first prediction of a functional domain in the RNA polymerase.

It is suggested that a kind of parallel evolution governs the polymerases in their diversification processes. DNA-directed DNA polymerase, reverse transcriptase and RNA-directed RNA polymerase have both synthesis and cleavage activities of nucleotide polymer. The cleavage activity in DNA polymerase or reverse transcriptase is carried by single globular domain. It suggests that the polymerases, including DNA-directed RNA polymerases, shear a pattern of molecular evolution in which a common ancestor of polymerase has diverged by capturing various nucleases as the functional domains.

In the analysis of the models of the domain, it is shown that several amino acid residues of the models are mechanically unstable and have improper solvent accessibility. The residues are hydrophobic/ambivalent amino acids which are exposed to the domain surface. Since the domain is a part of the large subunit of the RNA polymerase, it is probable that the residues

take part in the interface between the domain and other regions of the protein. This observation brings a question on adaptive evolution accompanied by domain fusion.

A fusion of functional units of proteins would not be completed by a simple merging of two proteins but it requires adaptive replacements of amino acids which are necessary for the stabilization of fused conformation or coupling of the functions. The pattern and extent of adaptive replacements is studied by using reverse transcriptase and RNase HI.

Reverse transcriptase is composed of polymerase, connection and ribonuclease H domains. RNase HI is a homolog of the RNase H domain, though it exists as a single domain enzyme. Amino acid substitution patterns between the free (RNase HI) and the integrated (RNase H domain) forms of domain were compared at each residue site. The results show that 8 sites have drastic substitution pattern between the two forms. The spatial distribution of the 8 sites reveals that only 4 sites are involved in the domain interface. The substitution pattern of the 4 sites is change from hydrophilic to hydrophobic or ambivalent amino acids. The residues make hydrophobic interactions or improve the fitting of domain surfaces at the interface.

The number of residue sites involved in the interface on RNase H domain is 29. The 4 adaptive sites are only 14% of them. It suggests the essential part of the adaptation of the fused domain was completed by a few residues. It might explain why the strategy of recruiting nucleases has been independently accepted by various polymerases; reuse of existing nucleases is easier than a scratch-build and adaptation of the recruited domains can be done by a few amino acid replacements.

Chapter 1

Introduction

1.1 Two different processes in protein evolution

There are two major fundamental processes in protein evolution. One is substitution of amino acid residues. The other is adding or removing blocks of amino acid residues such as domains or modules. Similarity in three-dimensional (3D) structures or amino acid sequences is frequently found among non-orthologous proteins (Fig.1-1a). It implies that gene duplications and following modification of the genes are responsible for production of variety of proteins. However, it is often the case that only a part of protein shows similarity to a part of other protein (Fig. 1-1b). It is also often the case that distantly related proteins show large differences in their numbers of amino acid residues or local 3D structures. It suggests that not only amino acid substitutions but also insertions and deletions of peptide blocks play an important role in functional divergence of proteins.

Both of the two processes can change sequences and 3D structures of proteins. However, they could be distinguished in their roles in protein evolution. Substitutions change a few (only one in most cases) amino acid residues at a time and they happen in very long time intervals. Typically, substitution rate is 10^{-9} at one residue site in a year (Kimura 1983). However, a block insertion/deletion introduces or removes many amino acid residues in one event (Fig.1-2).

When these two processes are viewed from protein 3D structures, the difference seems to be more significant. Proteins have their specific 3D structures which are essential for their functions. The roles of amino acid residues of a protein in function and 3D structure formation are different. There are some functionally or structurally essential amino acid residues, which are quite conservative in protein evolution. On the other hand, there are less important residues, which are not so conservative

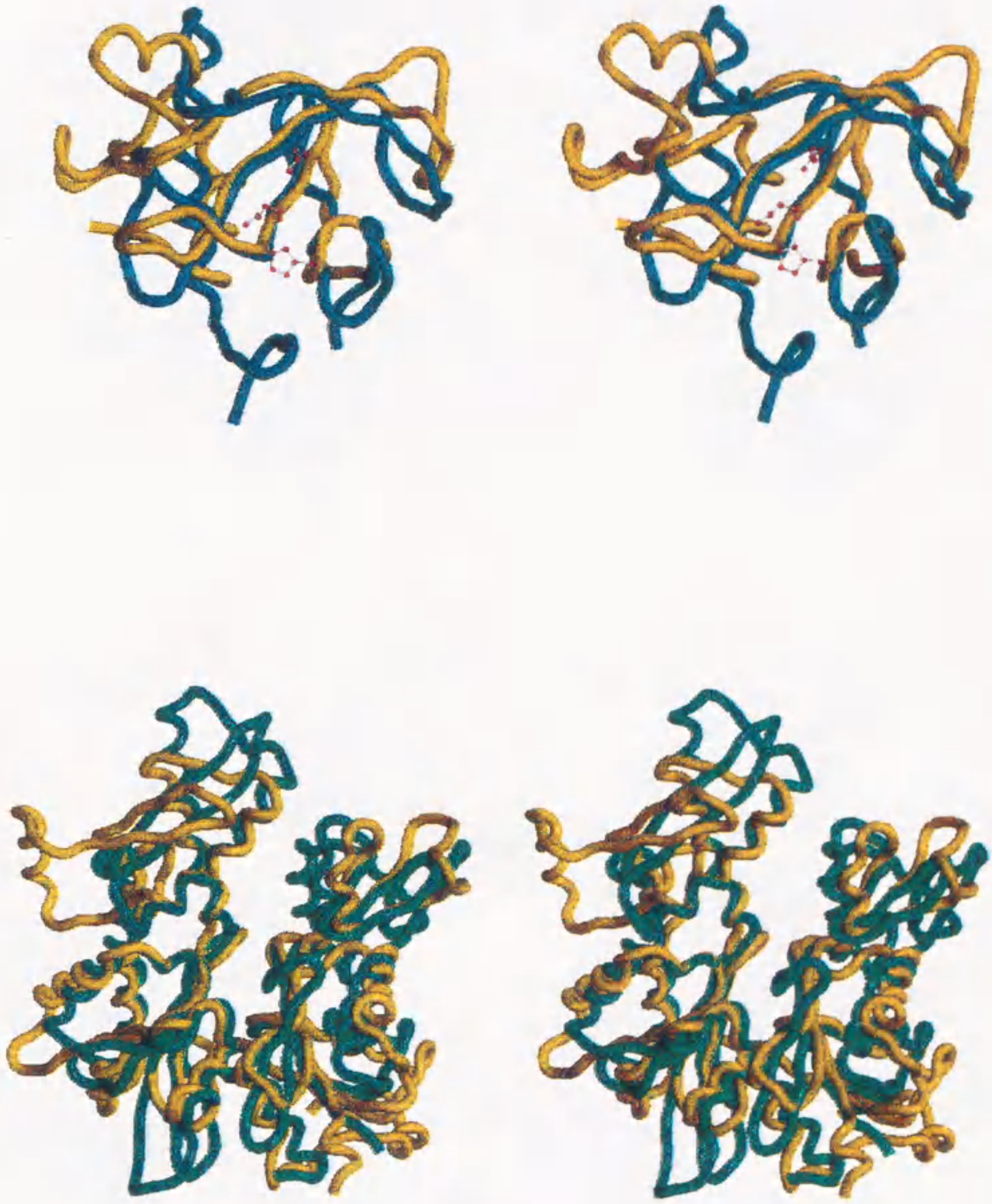


Figure 1-1a (See page 6 for legend)

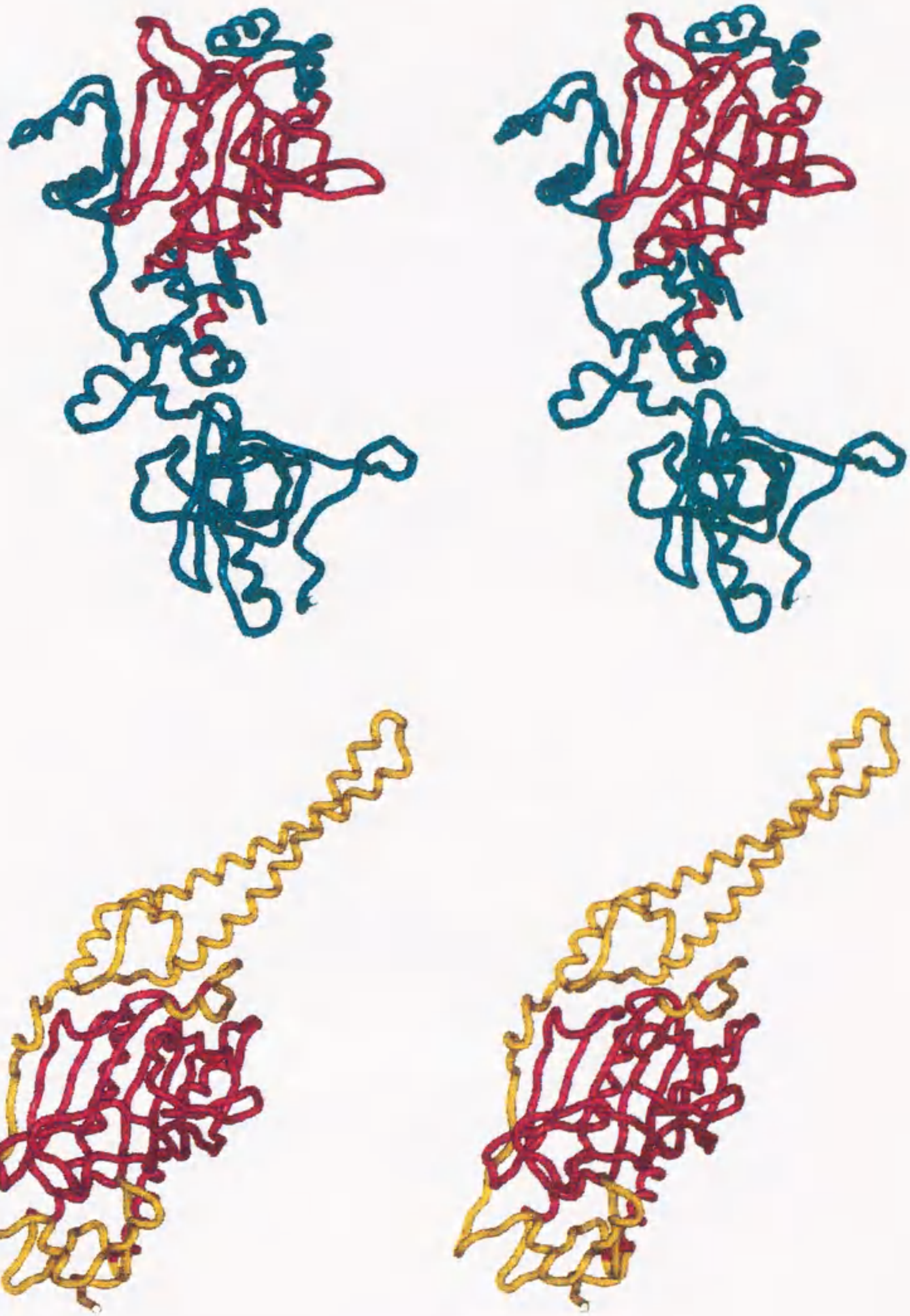


Figure 1-1b (See page 6 for legend)

Figure. 1-1. ***a.*** Homologous proteins are well conserved in their 3D structures, even the sequence similarity is low. Top; The stereo pair of superimposed backbone structures of barnase and RNase T1 in tube models. The former is presented in yellow and the latter in blue. They are both microbial ribonucleases; barnase is from *Bacillus amyloliquefacience* and RNase T1 is from *Aspergillus oryzae*. Though the sequence identity is only about 15% between them, the 3D structures are conserved, especially around the β sheets which compose the catalytic centers. The catalytic sites are composed of Glu, Arg and His residues in both enzymes. The catalytic sites of barnase are shown in ball and stick models (Hill *et al*. 1983). Bottom; the stereo pair of the backbone structures of actin and heat shock protein 70 (Hsp70) in tube models. Actin is shown in yellow and Hsp70 in green. Though the amino acid sequence similarity between the two proteins is also only trace level, the conformations are conserved. ***b.*** Structural diversity of homologous proteins is often generated by insertion/deletion of domains or modules. The back bone structures of the monomer of Aspartyl-tRNA synthetase from *Saccharomyces cereviciae* (top) and seryl-tRNA synthetase from *Thermus thermophilus* (bottom) are presented. They belong to class II aminoacyl-tRNA synthetases (ARS). The class II ARSs shear the domain which binds ATP. The ATP binding domains are colored in red in both structures. The domains of the ARSs which bind and recognize their cognate tRNAs or amino acids are largely different in the sequences, 3D structures and their mutual locations (in both primary and tertiary structures). These specific domains are colored in blue and yellow in Asp-tRNA synthetase and Ser-tRNA synthetase, respectively. Most likely, additions of these specific domains to a common ATP binding domain have led to their differentiated functions (Cusak 1995).

conservative. It seems that the requirement for a protein to be active is proper distribution of some essential residues in space. Amino acid substitutions remove or introduce only one residue in a event, while insertion/deletion can insert or delete a set of residues in one event. An insertion of blocks, which are domains or modules, can introduce entire residues required for a function at a time. It would be more important that insertions bring amino acid residues at previously unoccupied space. This is quite important for creation of a new function of proteins. The evolutionary processes of protein 3D structures by reorganization of protein building blocks are largely open questions in the study of molecular evolution.

1.2 Nuclease domains in nucleotide polymerases

The study of molecular evolution of polymerases by insertion of functional domain has been initiated by the finding of RNase-like domain in DNA-directed RNA polymerase II β subunit (Shirai and Go 1991). The RNase-like domain shows weak amino acid sequence similarity with bacterial ribonucleases such as barnase and RNase T_1. In spite of the low sequence identity, it was found that the active sites and

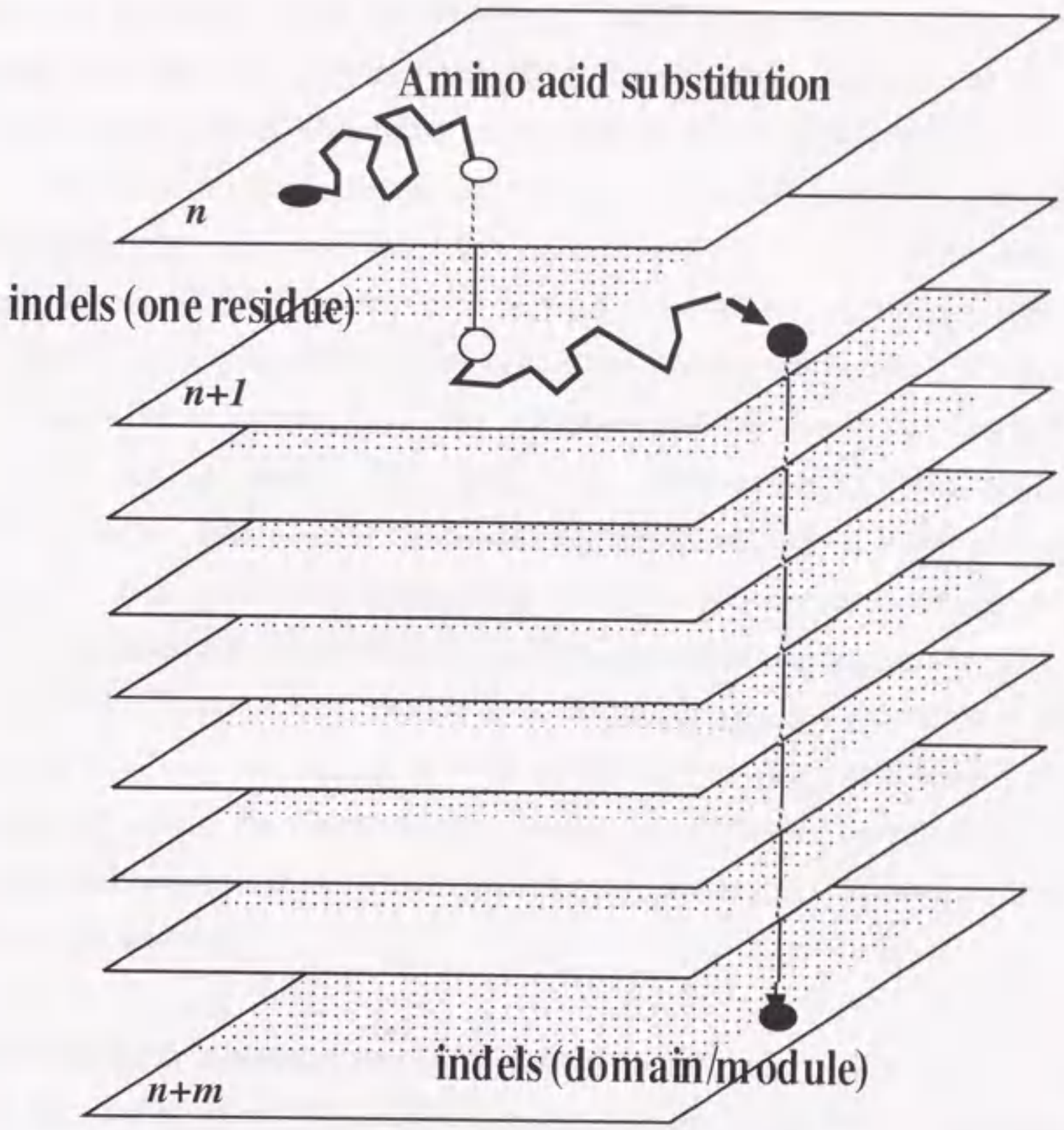


Figure 1-2. The difference between one amino acid residue substitution and insertion/deletion is schematically represented. Plates represent "sequence space". Two proteins composed of the same number of amino acid residues are on the same plate. Proteins drift within a plate by amino acid substitutions. They move to other plates by indels.

structurally important residues of RNases are conserved in the domain of the RNA polymerase. Possibly the RNase-like domain might work for proofreading activity in mRNA transcription like a DNase domain in DNA polymerase does in replication of DNA. The other putative function of the domain is degradation of nascent RNA in abortion of progressive transcription.

The RNA hydrolysis activity of the enzyme was not known at the point when the

RNase-like domain was found (Shirai and Go 1991). However, it was reported soon after the finding (Reines 1992). The RNA cleavage activity is thought to be involved in abortive initiation process and recovery of halted transcription complex. Active center of the RNA cleavage activity of the RNA polymerase is not known at this point. The RNase-like domain is a candidate for the catalytic center of the activity.

Not only the RNA polymerase but also many other nucleotide polymerases bear nucleotide cleavage activities (*e.g.* exo-DNase activity of DNA-directed DNA polymerases, RNase H activity of reverse transcriptases) (Kornberg 1980). The 3D structures of *E. coli* DNA polymerase I and reverse transcriptase of type 1 human immunodeficiency virus show that the portions which bear the nuclease activities conform globular domains (Ollis *et al.* 1985, Kohlstaedt *et al.* 1992). The RNase-like domain is the first case of functional domain prediction in RNA polymerases and suggests that the domain composition of RNA polymerase is similar to the other polymerases. The nuclease domains in the polymerases might give one typical example of evolutionary process by functional or structural units reorganization in protein. The finding of RNase-like domain in RNA polymerase implies that a scheme of molecular evolution govern the diversification process of nucleotide polymerases, in which a primordial polymerase has differentiated by recruiting the different nucleases as the functional domains.

1.3 The adaptive mechanisms of integrated domain

The finding of the RNase-like domain brings another problem on protein evolution by domain integration. How do structural/functional units introduced into other proteins adapt to the new molecular environment? Molecular adaptation of an integrated domain was studied on RNase H domain of reverse transcriptase (RT) (Shirai and Go 1996).

The RT consists of the N-terminal polymerase domain and the C-terminal RNase H domain. Type I RNase H (RNase HI) is a homolog of the RNase H domain though it exists in the form of single domain enzyme. The RNase HI is the free form of domain and the RNase H domain is the integrated form. The polymerase domain of RT shows similarity with the polymerase domain of *Esherichia coli* DNA polymerase I. It implies that integration of RNase H domain has occurred in the course of the RT evolution. Probably, a fusion of a primordial polymerase domain and a RNase HI's ancestor has generated the contemporary RTs.

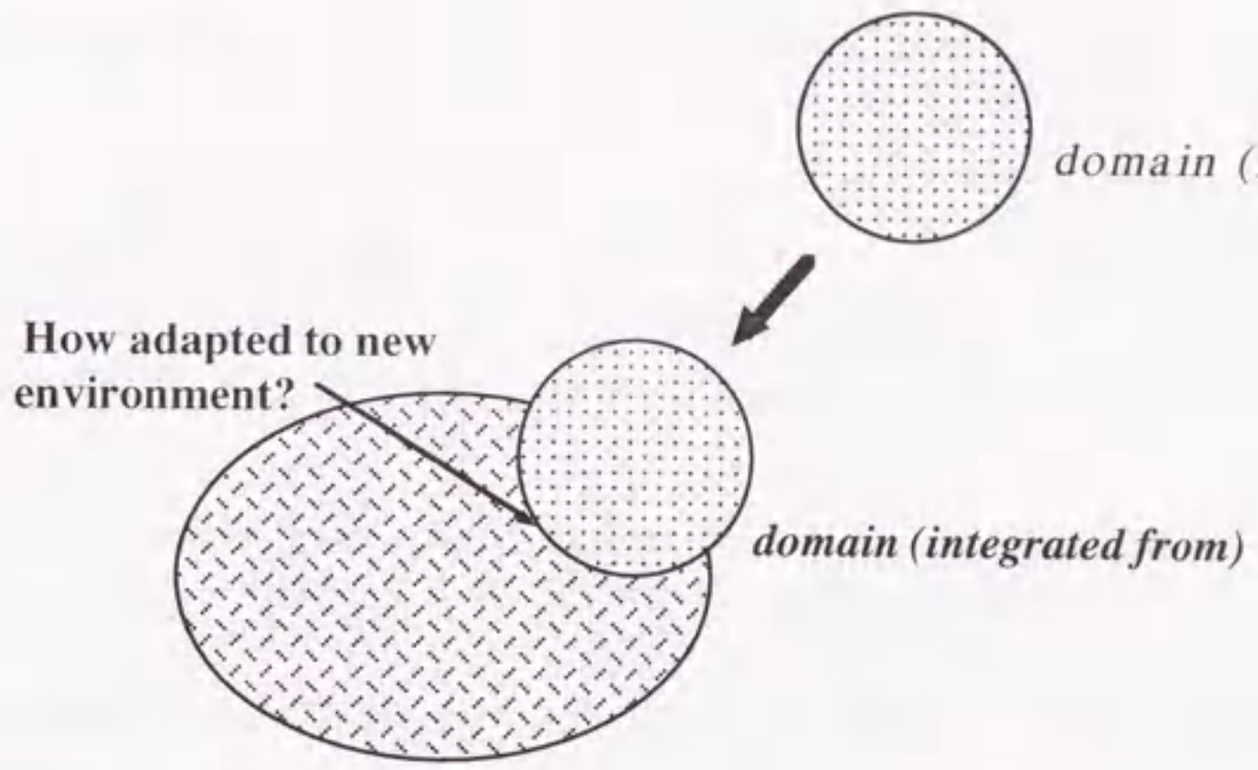


Figure 1-3 Adaptive evolution of integrated form of domain.

The RTs and the RNase HIs give good example of the integration event of functional units (Fig. 1-3), because the direct offspring of the integrated domain, RNase HI, exists and its 3D structure and several related amino acid sequences are known. The amino acid residue composition of each residue site of the integrated form (RNase H domain) and the free form (RNase HI) of domains were compared to extract the domain integration accompanied amino acid replacements. The results show that 8 residue sites have been substituted in drastic manner in the adaptation process.

The 3D structures of the domains show that 4 of the 8 residue sites are placed at the domain interface. They have been substituted from hydrophilic amino acids of RNase HI (free form) into hydrophobic or ambivalent ones in RNase H domain (integrated form). It seems that these replacements are responsible for stabilization of the fused conformation by hydrophobic interactions or the improved surface compatibility at domain interface.

The 4 adaptive sites are only 14% of 29 inter-domain contact sites in the integrated forms of the domain. The result suggests that an adaptation of integrated functional unit to molecular environment can be completed by a few amino acid replacements. The

suggestion might explain why the nucleotide polymerases have accepted the evolutionary strategy of nuclease domain capture in the course of functional differentiation.

Cusak, S. 1995. *Nature Struct. Biol.* 2: 824-831.

Delarule, M. D., Poch, O., Tordo, N., Moras, D. and Argos, P. 1990. *Prot. Engng.* 3: 461-467.

Hill, C., Dodson, G., Heinemann, U., Saenger, W., Mitsui, Y., Nakamura, K., Borisov, S., Tischenko, G., Polyakov, K. and Pavlovsky, S. 1983. *Trends Biochem. Sci.* 8: 364-369.

Kimura, M. 1983. in The Neutral Theory of Molecular Evolution. (Cambridge Univ. Press, Cambridge, England)

Kornberg, A. 1980. in DNA Replication (Freeman, San Fransisco). pp.101-166.

Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A. and Steitz, T. A. 1992. *Science* 256: 1783-1790.

Ollis, D. L., Brick, P., Hamlin, R. Xuong, N. G. and Steitz, T. A. 1985. *Nature* 313: 762-766.

Reines, D. 1992. *J. Biol. Chem.* 267: 3795-3800.

Shirai, T. and Go, M. 1991. *Proc. Natl. Acad. Sci. USA* 88: 9056-9060.

Shirai, T. and Go, M. 1996. *J. Mol. Evol.* (in press)

Chapter 2

RNase-like domain in DNA-directed RNA polymerase II

DNA-directed RNA polymerase is responsible for gene expression. Despite its importance, many details of its function and higher order structure still remain unknown. A local sequence similarity between the second largest subunit of RNA polymerase II and bacterial RNases Ba (barnase), Bi, and St is found. The most remarkable similarity is that the catalytic sites of the RNases are shared with the eukaryotic RNA polymerase II subunits of *Drosophila melanogaster* and *Saccharomyces cerevisiae*. Several amino acids conserved among the RNases and the RNase-like domains of the RNA polymerase subunits are located in the neighborhood of the catalytic sites of barnase, whose 3D structure has been resolved. 3D structural models of 4 RNase-like domains were constructed by using a method of homology modeling. The models were surveyed of their mechanical stability and solvent accessibility profiles of the residues. They are essentially stable during a 10ps molecular dynamics simulation. Solvent accessibility of the most of the model residues are proper values for a globular domain according to statistics on 28 fine structures of the globular proteins. These inspections support the conservation of the conformations between the domains and the RNases. The conservation of the structure and the active sites suggests the functional importance of the RNase-like domain of the RNA polymerase subunits and indicates that the domain may have RNase activity. The location of the RNase-like domain relative to the region necessary for RNA polymerization is similar to the relative proximity of 5' → 3' or 3'→ 5' exo-nuclease and the region of polymerase activity of DNA polymerase I. The RNase-like domain might work in proofreading, as in RNA-directed RNA polymerase of influenza virus, or it may contribute to RNA binding through an unknown function.

2.1 Introduction

A DNA-directed RNA polymerase is one of the most complicated apparatus of enzymes which is responsible for gene expression. Prokaryotic RNA polymerases are fundamentally a mono-form complex of four major subunits (α , β , β' and σ)(Yura and Ishihama 1979), while eukaryotic RNA polymerases consist of three distinct forms (RNA polymerases I, II and III) which show sequence homology. These eukaryotic polymerases are each made up of two large subunits and more than eight smaller components (Chambon 1975). In spite of these differences, it has been previously demonstrated that two large subunits of prokaryotic and three forms of eukaryotic RNA polymerases have local amino acid sequence homology. The amino acid sequences of the prokaryotic β' subunits and the largest subunits of eukaryotic RNA polymerases show extensive similarity in six regions (Allison *et al.* 1985). Similarly, the prokaryotic β subunits and the eukaryotic second largest subunits show local sequence similarity over nine regions, A to I (Fig.2-1). These facts suggest a monophyletic descent of RNA polymerases and functional similarity of the corresponding subunits of prokaryotes and eukaryotes (Falkenburg *et al.* 1987).

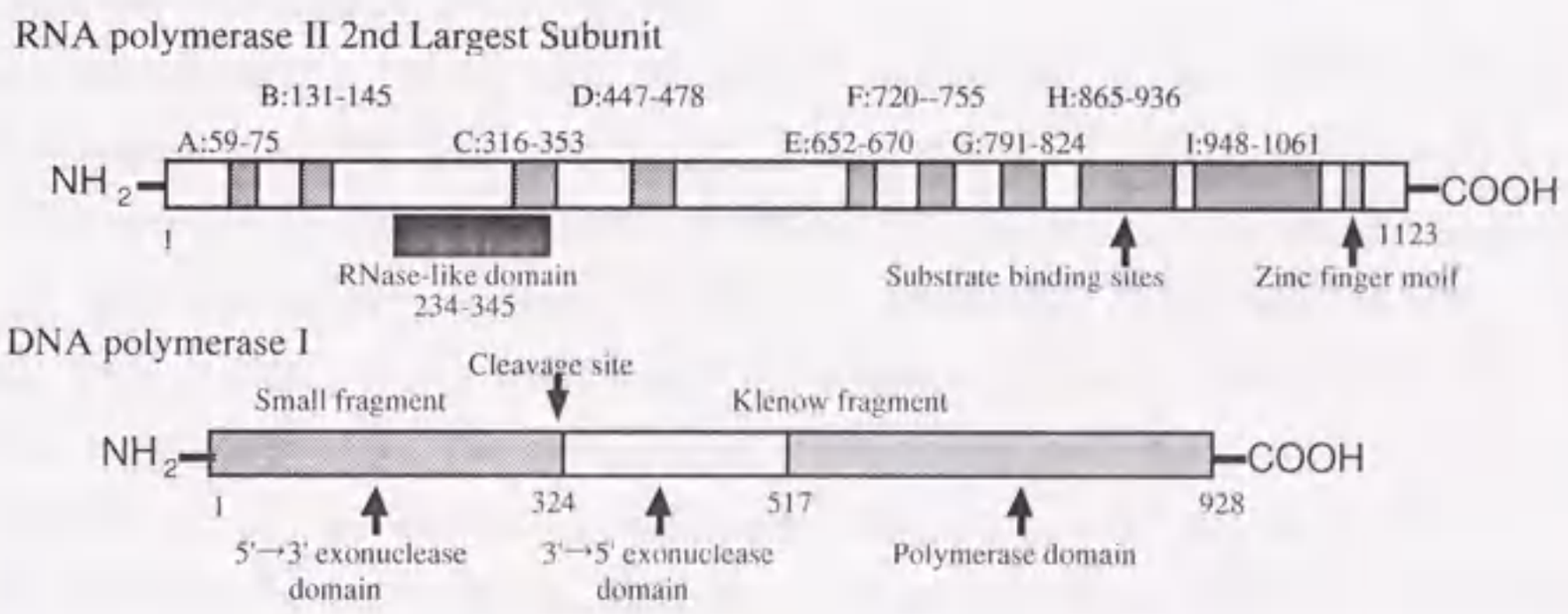


Figure2-1 Schematic diagrams of the functional regions of the second largest subunit of *D. melanogaster* RNA polymerase II and those of DNA polymerase I of *Escherichia coli* (Komberg 1980, Ollis *et al.* 1985). In the RNA polymerase subunit, the region having sequence similarity with barnase and the other bacterial RNases is indicated by filled bar. The homologous regions (A to I) of this subunit and the prokaryotic RNA polymerase subunits are also shown (Falkenburg *et al.* 1987, Sweetser *et al.* 1987).

Genetic and biochemical studies indicate that the β' subunit of prokaryotic RNA polymerase binds DNA, and that the β subunit binds the nucleotide triphosphate substrate, interacts with the σ subunit, and contributes to RNA catalytic function (Yura and Ishihama 1979). It has been suggested that the homologous counterparts of eukaryotic RNA polymerase subunits have similar functions. Some functional sites have been investigated using the affinity labeling technique. Substrate binding sites have been mapped to Lys or His residues in the homologous region H (Fig.2-1) of the second largest (β -like) subunit of eukaryotic RNA polymerase II (Riva *et al.* 1990) and those in the β subunit of prokaryotes were mapped to the corresponding region (Grachev *et al.* 1989). However, in spite of their functional importance, the active sites of each subunit in the transcription process still remain largely unknown. The existence of the multiple conserved regions suggests that any function might be associated with these regions.

A part of the sequence of the second largest subunit of RNA polymerase II of *D. melanogaster* (Falkenburg *et al.* 1987) and the subunit of *S. cerevisiae* (Sweetser *et al.* 1987) show weak sequence similarity with the bacterial RNases including barnase (Hartley and Barker 1972). The catalytic sites, Glu, Arg and His, of barnase and other microbial RNases were exactly aligned with Glu, Arg and His, respectively, in the RNA polymerase subunits. Hydrophobic residues conserved among the RNases and the RNase-like domains of the RNA polymerase subunits are located close to the catalytic sites and help to form the hydrophobic core of barnase (Mauguen *et al.* 1982). Based on the sequence similarity of the proteins, 3D atomic models of 4 RNase-like domains were constructed by a method of homology modeling. The constructed models are the domains of *D. melanogaster*, *S. cerevisiae*, *Shizomyces pombe* and *Homo sapiens*. The models were tested for their mechanical stability by molecular dynamics computer simulations *in vacuo* at 300K for 10ps. The models were stable during the simulation. The models were also tested by solvent accessibility profiles of the side chains of the amino acid residues. The analysis is based on the statistics of side chain solvent accessibility of 28 fine and high-resolution crystal structures of globular proteins. It is shown that the most of model side chains have proper accessibilities. However, some of them are significantly deviated from the average. The most of the aberrant residues are hydrophobic amino acids which are exposed on the surface of model structures. These residues would be acceptable, because there may be some residues in the RNase-like domain which make contact with the other portions of the 2nd largest subunit of RNA polymerases.

The analysis of the amino acid sequences and constructed 3D models suggested the

similarity of these two proteins are significant. It is speculated that the corresponding portion of the eukaryotic RNA polymerases might have RNA cleavage activity. It is possible that such activity is regulated by other regions or subunits of RNA polymerase.

2.2 Amino acid sequence analysis of RNase-like domains

Barnase is an endo-ribonuclease produced by *Bacillus amyloliquefaciens* which hydrolyzes a single strand RNA at a 3' phosphodiester bond of ribonucleotide residues with weak specificity. It recognizes purines better than pyrimidines, and prefers guanine to adenine. Catalytic sites of barnase and other microbial RNases have been identified as Glu-73, Arg-87 and His-102 in the numbering system of barnase (Hill *et al.* 1983). Sequence similarity between barnase and the RNA polymerase subunit of *D. melanogaster* was found in the Swissprot data base. According to some previous studies, RNases Bi and St were aligned with barnase (Hill *et al.* 1983), and the RNA polymerase subunit of *S. cerevisiae* was aligned with that of *D. melanogaster* (Berghofer *et al.* 1988). In order to refine the alignment, 20 amino acids were classified into five groups according to their substitution behavior during molecular evolution. By referring to the three-dimensional structure of barnase, insofar as was possible, insertion/deletion sites were located on the surface.

The regions of the *S. cerevisiae* and *D. melanogaster* RNA polymerase subunits, having sequence similarity with barnase (Fig.2-2), involve almost the whole of C-region, one of the nine homologous regions between the second largest subunits of eukaryotic RNA polymerase IIs and the corresponding subunits of prokaryotes (Fig.2-1). C-region, consisting of about 40 amino acid residues, corresponds to a carboxyl-terminal part which composes about one-third of the RNase-like domain of the RNA polymerases.

Hydrophobic interactions stabilize the core region of a protein (Chothia and Lesk 1987) and the three-dimensional structure of a protein is more stable evolutionary than its amino acid sequence (Kimura 1983). Thus, the tertiary structure should be useful for investigating distantly related molecules. Using the information of the X-ray crystallographic structure of barnase (Mauguen *et al.* 1982), the spatial arrangement of the conserved residues between the RNases and the RNA polymerase subunits were examined. A proposed model of the RNase-like domain of the RNA polymerase II subunit based on its sequence similarity with barnase was used to confirm the potentiality

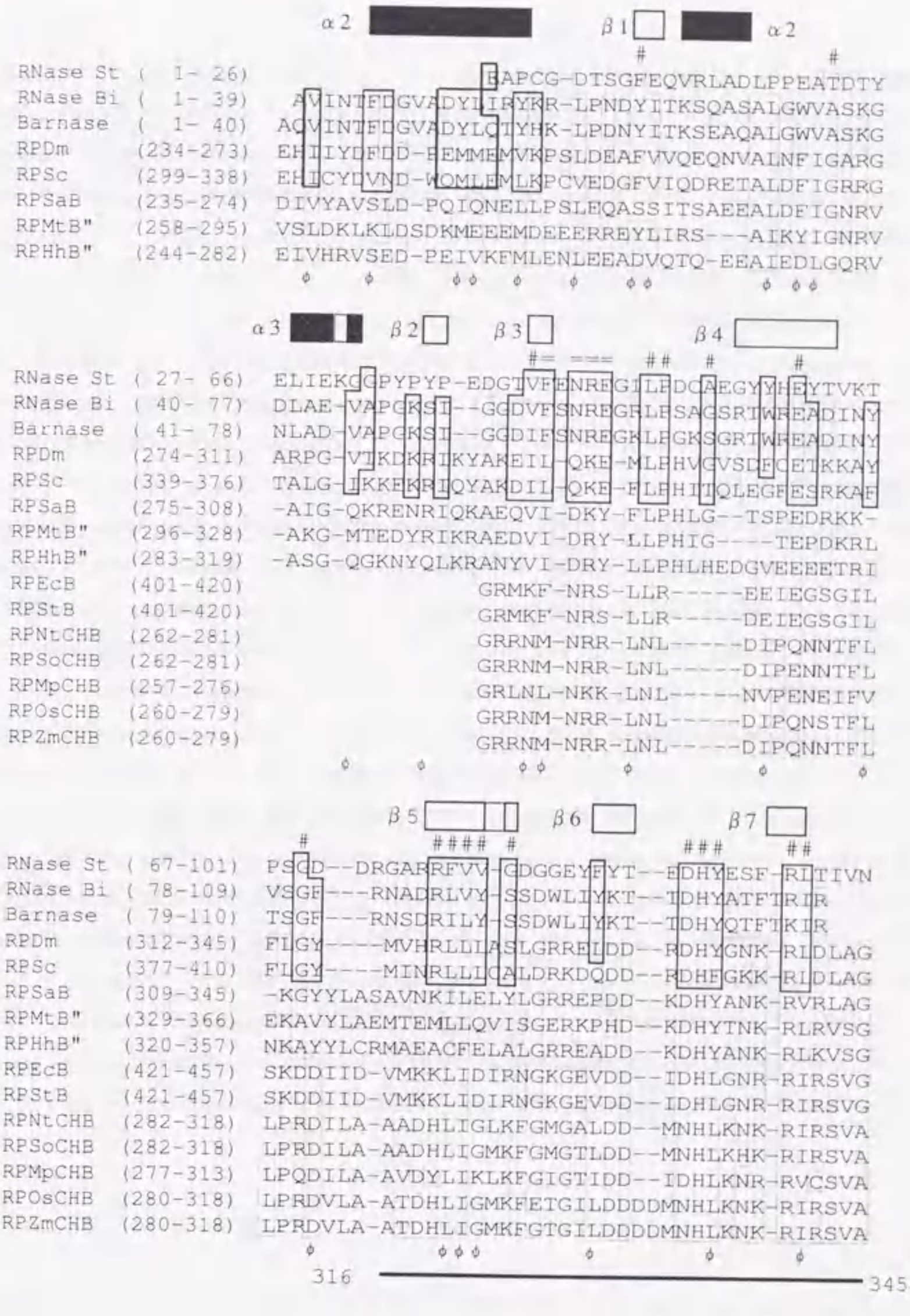


Figure 2-2 (See next page for legend)

Figure 2-2. Amino acid sequence alignment of the bacterial RNases and the RNA polymerase subunits. Prokaryotic RNases; *Streptomyces erythreus* (RNase St) (Yoshida *et al.* 1976), *Bacillus intermedius* (RNase Bi) (Aphanasenko *et al.* 1979), *B. amyloliquefaciens* (barnase) (Hartley and Barker 1972). The second largest subunits of eukaryotic RNA polymerase II; *D. melanogaster* (RPDm) (Falkenburg *et al.* 1987), *S. cerevisiae* (RPSc) (Sweetser *et al.* 1987). The archaebacterial RNA polymerase B subunit; *Sulfolobus acidocaldarius* (RPSaB) (Puhler *et al.* 1989). The archaebacterial RNA polymerase B" subunits; *Methanobacterium thermoautotrophicum* (RPMtB") (Berghofer *et al.* 1988), *Halobacterium halobium* (RPHhB") (Leffers *et al.* 1989). Eubacterial RNA polymerase β subunits; *E.coli* (RPEcB) (Ovchinnikov *et al.* 1981), *Salmonella typhimurium* (RPStB) (Listisyn *et al.* 1988). Chloroplast RNA polymerase β subunits; tobacco (*Nicotiana tabacum*) (RPNtCHB) (Ohme *et al.* 1986), spinach (*Spinacia oleracea*) (RPSoCHB) (Hudson *et al.* 1988), liverwort (*Marchantia polymorpha*) (RPMpCHB) (Umesono *et al.* 1988), rice (*Oryza sativa*) (RPOsCHB) (Hiratsuka *et al.* 1989), maize (*Zea mays*) (RPZmCHB) (Hu and Bogorad 1990). Amino acid residues are indicated by a one-letter code and gaps by a hyphen. Sites are boxed by thin lines when occupied with identical or similar residues (G=A=S=T, V=L=I=M=F=Y=W, N=Q=D=E, R=K=H) of more than 80% when 8 sequences are aligned or 50% when 15 sequences are aligned. Sites of the RNases and the eukaryotic RNA polymerase subunits are boxed by thick lines when occupied with identical or similar amino acid residues in at least four sequences. Bold letters indicate the three catalytic sites of RNases and their corresponding amino acid residues of RNA polymerase subunits in the alignment. The putative substrate binding sites of RNases are indicated by = above the sequences. Numbers in parentheses on the left indicate the amino acid residue numbers. The C-region of the RNA polymerase subunits is underlined with the residue numbers of the fruit fly subunit. The secondary structures of barnase are represented at the very top of each row. The sites conserved in all three of the RNases and the two eukaryotic RNA polymerase subunits are indicated by # above the sequences. The symbol ϕ under the sequences indicates the conserved sites which are occupied by the hydrophobic residues (V, L, I, M, F, Y and W) in at least four of the RNases and in the eukaryotic RNA polymerase subunits.

for the involvement of the conserved residues in the functional and structural importance of the domain. It was examined whether the conserved residues were located close to the catalytic region or whether they contributed to form the hydrophobic core of the RNase-like domain.

2.3 Homology modeling of the RNase-like domains

Based on the sequence similarity, three-dimensional models of the RNase-like domains were constructed by method of homology modeling. In spite of low sequence similarity, 3D structures are usually conserved between homologous proteins. When amino acid residues of one protein are placed on the corresponding positions of 3D structure of homologous proteins, the spatial distribution and solvent accessibility profiles of modeled residues will take the proper values for the residue types. Thus, the homology models can be used for an analysis of significance of sequence similarity.

Atomic models were constructed for 4 RNase-like domains of eukaryotic RNA polymerase II. They are from *D. melanogaster* (Falkenberg *et al.* 1987), *S. cerevisiae* (Sweetser *et al.* 1987), *S. pombe* (Kawagishi *et al.* 1993) and *H. sapiens* (Acker *et al.* 1992). The 4 models were used to extract their consensus features. The prototype of the models is the crystal structure of barnase (Baudet and Janin 1991).

The modeling was preformed with the program BIOGRAF (Mayo *et al.* 1990) by the following protocols for each model.

Step 1: The 3D structure of barnase was energy minimized until root mean square force (RMSF) became less than 0.1kcal/mol/Å.

Step 2: Two insertions and four deletions were introduced into the minimized barnase structure. Insertions/deletions are one of the most annoying problems in homology modeling, especially when there are long insertions in target sequence. This is because there is no reliable method to deduce the conformation of long insertions without any templates. In the case of the RNase-like domain, the two insertions are composed of only 1 and 2 residues. There is no indel among 4 target sequences. So the problem is less significant in this case. The deletions were introduced by omitting the atoms of deleted residues from barnase structure. The residues at the both ends to the deleted residue(s) were connected by a peptide bond of aberrant bond length and angle. The insertions were introduced by cutting the peptide bonds of the insertion sites. Two peptide bonds were made between C-terminal of the inserted residue and the C-terminal residue of the insertion site, and between N-terminal of the inserted residue and the N-terminal residue of the insertion site. The former peptide bond was made by the proper bond length and angle, while latter was made by aberrant ones.

Step 3: The aberrant bond geometries were corrected by an energy minimization of the structure. During the course of the minimization, peptide bond dihedral angles (ω) of the

inserted residues and those within 3 residues from the inserted and deleted residues were constrained to 180° to avoid *cis* conformation. The constraint coefficient was 200kcal/mol/Å^2 for each bond. The convergent condition was the same as in step 1 (RMSF < 0.1kcal/mol/Å)

Step 4: Side chains of the residues which are different from the corresponding residues of barnase were replaced with those of the RNase-like domain. Initially, the conformations of the side chains of the residues were set to canonical ones. Later, the conformations were corrected manually to avoid steric hindrance, by referring the side chain conformations of barnase.

Step 5: Steric hindrance in the models was relaxed by energy minimization of the model conformation. The condition for convergence was same as in step1. To detect local steric hindrance of the models, inter-atomic potential energy, covalent bond energy, and dihedral angle energy were calculated for every possible set of atoms. It was confirmed that each inter-atomic energies of non-bonded atoms was less than 20kcal/mol, and all inter-atomic energies of bonded atoms, covalent bond energies and dihedral angle energies were less than 10kcal/mol.

2.4 Molecular dynamics analysis of the models

The obtained model structures were tested for their mechanical stabilities. Unstable conformations of proteins (*e.g.* with steric hindrance, lack of hydrogen bonds) can be detected by molecular dynamics simulations. Molecular dynamics simulation of the 4 RNase-like domain models and energy minimized barnase X-ray structure were performed with AMBER3 (Weiner *et al.* 1984) *in vacuo* at 300K for 10ps. The distance dependent dielectric constant was employed and the electrostatic interactions were cut off at 8Å. The trajectories of the simulation were stored every 10 step (0.1ps). The average coordinates and the standard dislocation (standard deviation of distances from the average coordinates) of atoms were calculated from the 100 stored trajectories. The flexibility of a residue is represented by the average of standard dislocation of the atoms of the residue.

2.5 Solvent accessibility of side chains in model structures

Amino acid residues on protein 3D structures show apparent tendency for spatial

distribution depending on the physico-chemical properties of their side chains. Solvent accessibility is one of such properties which show significant difference depending on the hydrophobicity of side chains. The solvent accessibility of a residue is the fraction of accessible surface area of side chain atoms (in native conformation) in its maximum. Since the accessibility is normalized by the maximum surface area of the amino acid residue, it is comparable between different amino acids (Go and Miyazawa 1979).

Table II-I Crystal structures used in the solvent accessibility statistics

Name	BNL code	Resolution	Length(res.)	R-factor	Refference and Note
Phospholipase A2	1BP2	1.70	123	0.17	Dijkstra *et al.* 1981
Flavodoxin	2FCR	1.80	173	0.17	Fukuyama *et al.* 1992
Dihydrofolate reductase	3DFR	1.70	162	0.15	Bolin *et al.* 1982
Crambin	1CRN	1.50	46	0.11	Teeter 1984
Parvalbumin	1PAL	1.65	108	0.20	Declercq *et al.* 1991
Erabtoxin	3EBX	1.40	62	0.18	Smith *et al.* 1988
Ribonuclease T1	9RNT	1.50	104	0.18	Martinez-Oyanedel *et al.* 1991
Fattyacid binding protein	1IFB	1.96	131	0.19	Sacchettini *et al.* 1989
Lysozyme	2LZT	1.97	129	0.12	Ramanadham *et al.* 1987
Alpha-lytic protease	2ALP	1.70	198	0.13	Fujinaga *et al.* 1985
Elastase	6EST	1.80	240	0.20	Li de la Sierra *et al.* 1990
Plastocyanin	3PCY	1.90	99	0.16	Church *et al.* 1986
Pancreatic trypsin inhibitor	5PTI	1.00	58	0.20	Van Mierlo *et al.* 1991
FK506 binding protein	1FKF	1.70	107	0.17	Van Duyne *et al.* 1991
Azurin	2AZA	1.80	129	0.16	Baker 1988
Interleukin-1 β	4I1B	2.00	151	0.19	Veerapandian *et al.* 1992; N-term. 2res. absent
β -lactamase	3BLM	2.00	257	0.16	Herzberg 1991
L-arabinose-binding protein	7ABP	1.67	305	0.16	Vermersch *et al.* 1991; N-term. 1res. absent
Subtilisin	2ST1	1.80	275	0.14	Bott *et al.* 1988
Thermolysin	6TMN	1.60	316	0.17	Tronrud *et al.* 1987
T4 endnuclease V	1END	1.60	137	0.20	Morikawa *et al.* 1992; N-term .1res. absent
Ribonuclease HI	2RN2	1.48	155	0.20	Katayanagi *et al.* 1992
Rubredoxin	8RXN	1.00	52	0.15	Dauter *et al.* 1992
H-ras P21 protein	321P	1.40	166	0.21	Pai *et al.* 1990
Macromomycin	2MCM	1.50	112	0.16	Van Roey and Beerman 1989
CheY	3CHY	1.70	128	0.15	Volz and Matsumura 1991; N-term. 1res. absent
Ribonuclease A	7RAT	1.50	124	0.16	Tilton Jr. *et al.* 1992
Guanylate kinase	1GKY	2.00	186	0.17	Stehle and Schulz 1992

The solvent accessibility profiles of the residues in the models were calculated. They were compared with the average values which were obtained from globular proteins. To obtain the average and the standard deviation of the accessibility of each amino acid, twenty-eight crystal structures were surveyed (Table II-I). The selected structures are those which were determined to high-resolution (less than 2Å) with R-factors lower than 0.20. The solvent accessibility of 3853 residues (except for glycine residues) were collected to calculate the averages and the standard deviations.

2.6 RNase-like domain of eukaryotic RNA polymerases

All three of the catalytic sites of RNases (Hill *et al.* 1983) are conserved in the RNA polymerase subunits of *D. melanogaster* and *S. cerevisiae* (Fig.2-2). Moreover, the amino acid residues around the catalytic sites, Arg-87 and His-102, are conserved. The amino acid residues, Phe-56, Asn-58, Arg-59 and Glu-60 of barnase are important for substrate binding (Hill *et al.* 1983). These sites are aligned with the similar amino acid residues in the eukaryotic RNA polymerase subunits (Fig. 2-2).

As the divergence time between two molecules increases, the amino acid sequence similarity decreases. RNase St and barnase show only 22% sequence similarity. RNases Bi and barnase show a range of 17-23% sequence similarity to the RNA polymerase subunits from *D. melanogaster* and *S. cerevisiae*. Similarly, barnase shows 15% similarity with RNase T_1 from *Aspergillus oryzae*. However, the 3D structures of these molecules are well conserved, showing that they were derived from a common ancestor (Hill *et al.* 1983). The three catalytic sites are conserved among these microbial RNases. Uniquely conserved catalytic residues among the RNases, including the RNase-like domains of eukaryotic RNA polymerases, suggest that the RNase-like domain might have an essential function in RNA catalysis, a function which may include as yet undiscovered RNA cleavage activity.

2.7 Conserved Amino Acid Residues of RNase-like Domain

There are 22 sites which are occupied by identical or similar residues in all three of the RNases and in the RNase-like domain of the two eukaryotic RNA polymerase subunits (Fig.2-2). First, the conserved residues among the five sequences were examined on the tertiary structure of barnase. As mentioned above, seven of the sites are the catalytic or substrate binding sites of the RNases. Eleven of the remaining 15 sites are located near the catalytic center of barnase (Fig.2-3a), implying that all or some of these residues are

probably involved in the catalytic function of the RNase-like domain. Four residues, which are not located near the catalytic center, Tyr-24, Ala-37, Ser-67 and Gly-81 of barnase, correspond to Phe-257, Gly-270, Gly-300 and Gly-314 of the second largest subunit of *D. melanogaster*, respectively. In the case of the subunit of *S. cerevisiae*, these residues correspond to Phe-322, Gly-335, Thr-365 and Gly-379, respectively. The first site is occupied in the two subunits by Tyr or Phe which are both involved in the hydrophobic cores. The last three sites are occupied by small residues, Ala, Gly, Ser or Thr. These four residues often appear at β-turns and carry a critical role in protein folding (Wilmot and Thornton 1988). In fact, Ala-37 and Ser-67 of barnase are located on β-turns and Gly-81 makes a similar turn.

In at least four of the five sequences, 25 of the 113 aligned sites are occupied by the hydrophobic residues (Val, Leu, Ile, Met, Phe, Tyr and Trp) (Fig.2-2). Spatial distribution of the conserved hydrophobic sites were examined on barnase conformation. All of the 25 sites are occupied by the hydrophobic residues in barnase, and they constitute hydrophobic cores of barnase. Some of them are partially exposed on the surface of model structure to the extent similar to the exposure of some of barnase's hydrophobic sites (Fig.3b,c). This fact implies that the 25 hydrophobic sites make core regions in the RNase-like domain of RNA polymerase subunits similar to the core region of barnase. Since hydrophobic interaction is one of the most important factors in stabilizing protein conformation, the predicted hydrophobic cores in the RNase-like domain of the polymerase helps to confirm the obtained alignment among the RNases and the polymerase subunits and thus the model structure of the RNase-like domain.

Figure 2-3. (**a**) Conserved residues among the RNases and the RNase-like domains of the RNA polymerases are viewed on barnase conformation in stereo. The conserved residues among the three RNases and the two eukaryotic RNA polymerase subunits (indicated by # and = in Fig.2-2) are colored in magenta when they are located close to the catalytic center and in yellow when they are located at a distance from it. The catalytic sites of barnase, Glu-73, Arg-87 and His-102, are presented in a space filling model. (**b**) The hydrophobic surface patches of barnase and those of the predicted model conformation of the RNase-like domain. Left, the hydrophobic (yellow) and three catalytic residues (magenta) of barnase are shown. Right, the hydrophobic (yellow) and the catalytic sites of the RNases (magenta) conserved among the RNases and the RNase-like domains (indicated by ϕ and bold letters, respectively, in Fig.2-2) are shown on the model structure of the RNase-like domain, which is based on the structure of barnase. (**c**) Rear view of (b).

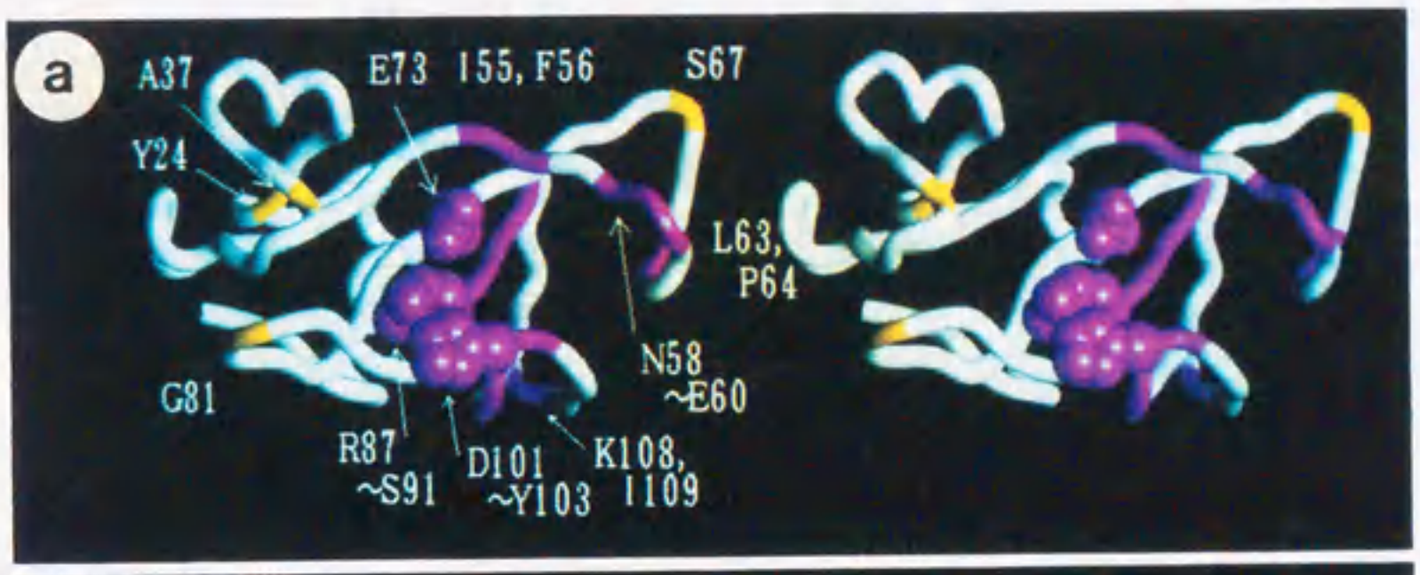


Figure 2-3 (See the previous page for legend)

2.8 Atomic models of the RNase-like domains

The atomic models of 4 RNase-like domains are constructed. Amino acid identity between each pair of the RNase-like domains is listed in Table II-II. Since they have accumulated a number of amino acid substitutions, constructed models show considerable diversity. So the consensus of these models would verify the reliability of the results. The models have similar back bone structures to barnase, and each other (Table II-III).

Table.II-II Sequence identity between RNase-like domains(%)

	D. melanogaster	*S. sereviciae*	*S. pombe*
H. sapiens	93.6	56.0	60.0
D. melanogaster		53.2	57.8
S. sereviciae			72.5

2.9 Mechanical stability of RNase-like domain models

The mechanical stability and profiles of solvent accessibility of the side chains were calculated for the models. Molecular dynamics simulations under the condition of no solvent molecule (*in vacuo*) at 300K were done. Protein conformations are stabilized by atomic interactions, such as hydrogen bonds, electrostatic, hydrophobic and *van der Waals* interactions.

Table II-III Comparison of RNase-like domain and barnase structures

	RMSD of superimposed C^{α} atoms (No. of atoms compared)			
	S. cereviciae	*S. pombe*	*D. melanogaster*	*H. sapiens*
barnase	1.26 (105)	1.26 (105)	1.31 (105)	1.42 (105)
S. cereviciae		0.65 (108)	0.87 (108)	0.96 (108)
S. pome			0.80 (108)	0.90 (108)
D. melanogaster				0.61 (108)

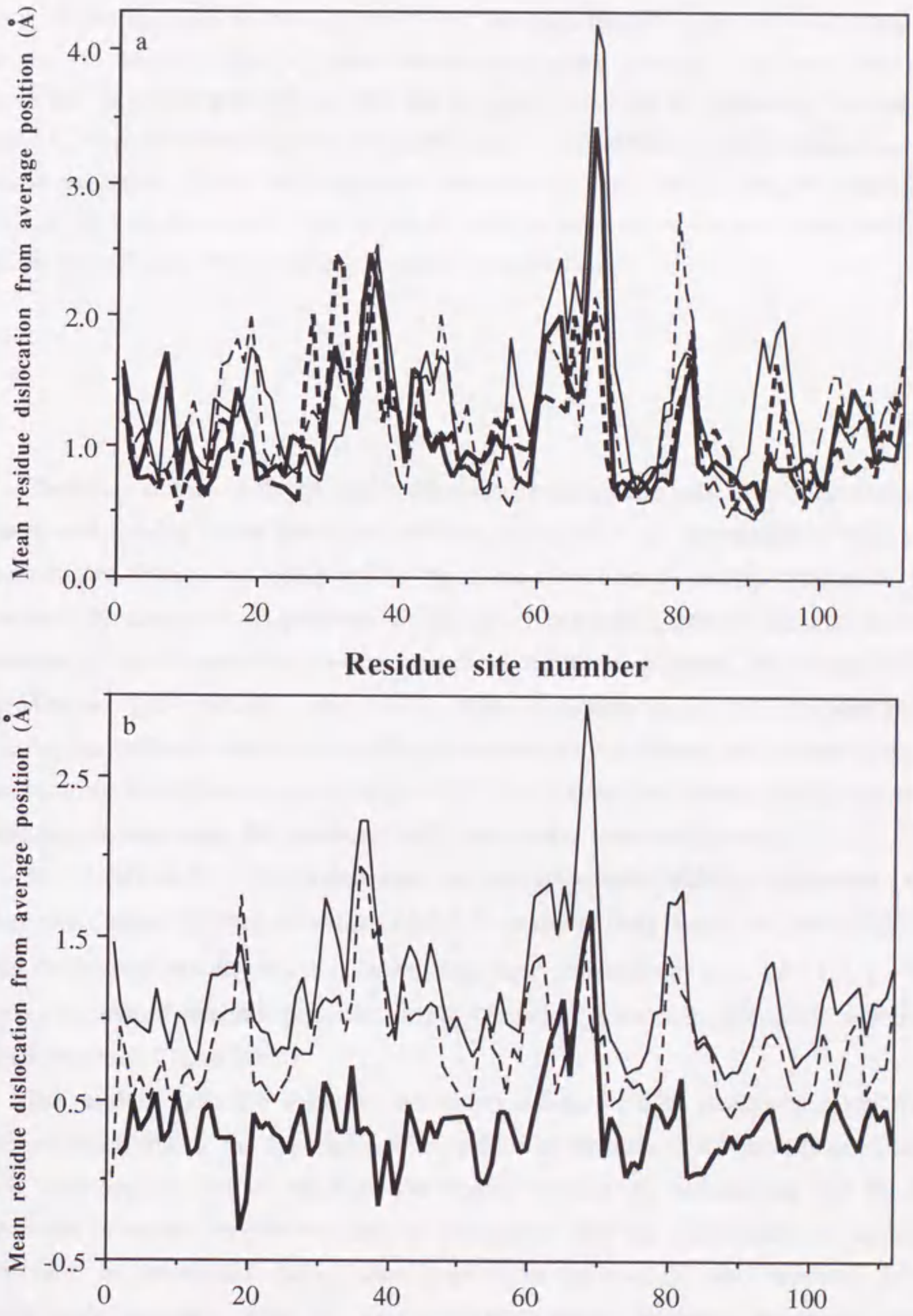


Figure 2-4 (See next page for legend)

Figure 2-4. Mean dislocation of residues from the average positions in the models of RNase-like domains and barnase during molecular dynamics simulations. **a.** The mean dislocations of the residues in the 4 RNase-like domain models are plotted against the residue numbers in the numbering system of barnase. The mean dislocation is averaged over the atoms of each side chain. The correspondence of curves and models are as the follows: thin line; *H. sapiens*, thick line; *D. melanogaster*, thin broken curve; *S. cerevisiae*, thick broken curve; *S. pombe*. **b.** The mean dislocation of the residues between the models and barnase. The thin line indicates the average dislocations of the 4 models. The middle line indicates the mean dislocation of barnase residues during a simulation in the same condition as the models. The difference of the two profiles is plotted by the thick line.

The traces of the simulation (trajectories) are translated into indicators of flexibility or mechanical stability of the structures. Mobility of residues are represented by their root mean square dislocations (RMSD) from the average positions during the simulation. The results of the simulation are presented in Fig.2-4a. The mobility profiles of corresponding residues of the 4 RNase-like domain models are similar each other. The average of the mobility over the 4 models is calculated to obtain the general profile of the models and it is compared with the dislocation profile of barnase structure during the simulation under the same condition for the models (Fig.2-4b). The profiles look similar, except for some local peptide segments. The residues, in the numbering system of barnase, 5, 8, 11, 32, 33, 36, 62-69 and 83 of the models have the average mobility which is lager than those of corresponding residues in barnase by 0.5Å or more. Only 4 residues, 64 and 66-68 have the average mobility which is larger than those of barnase by more than 1.0Å. The energy minimized final structures obtained by the 10ps dynamics are compared with their initial structures (Table II-IV).

The result of molecular dynamics simulation indicates that the models have kept their conformations during the simulation. The profiles of mobility of the models are similar each other and to that of barnase. The results confirm the assumption that the 3D structures of the RNase-like domains are compatible with the 3D structure of barnase. However, as mentioned above, some regions of the models have appeared to be significantly unstable than the corresponding parts of barnase. The cause of the perturbation will be discussed later with the result of the solvent accessibility profiles.

Table II-IV Comparison of atomic positions of the RNase-like domain models before and after the dynamics simulation

Model	No. of atoms	RMSD of atoms(Å)*	Maximum dislocation(Å)*
S. cerevisiae	1112	1.09	6.00
S. pombe	1105	1.06	4.83
D. melanogaster	1079	1.08	4.87
H. sapiens	1077	1.12	4.72

* between the initial and final models.

2.10 Solvent accessibility profile analysis of the RNase-like domain models

Polar amino acid residues are localized to surface of proteins, while apolar ones are clustered at the interior regions. Since the biased distribution is generally observed in globular proteins or globular domains, it can be used for a test of accuracy of model structures.

To obtain parameters for solvent accessibility analysis, 28 fine crystal structures were used. Since the accessibility is a fractional solvent accessible surface area, it takes values between 0 and 1. The accessibility is divided into 4 sections: 0.00~0.25, 0.25~0.50, 0.50~0.75 and 0.75~1.00. For each kind of amino acid residues, the sections are assigned with one of the three statuses, favored, neutral and disfavored, depending on the observed residue frequencies. When the number of one kind of the 20 amino acid residues found in a section is larger than the average by three folds of the standard deviation, the section is assigned as favored for the kind of the residue. If it is smaller than the average by three folds of standard deviation, it is assigned as disfavored. Otherwise, it is assigned as neutral. The average is defined as one fourth of total number of the residues of the kind of amino acid found in the sample proteins. The standard deviation is the square root of $(1\text{-}0.25)\times0.25\times$ {total number of residues of the kind of amino acid in sample proteins}. The results of the statistics are tabulated in Table II-V.

The solvent accessibility of side chains of the models are calculated and compared with the statistics. The residues, which fulfill the core regions and the area around the

catalytic center of barnase, are shown to have the favored solvent accessibility in the models (Figs. 2-5 and 2-6). The result supports the reliability of the models. Some of the model residues are found to have disfavored values (Figs. 2-5 and 2-7). Most of them are hydrophobic residues which are exposed on the surface of the model structures.

Table II-V Statistics of solvent accessibility of amino acids

Grou	Residues	Frequency in sections of accessibility				Total
		0.0 - 0.25	0.25 - 0.50	0.50 - 0.75	0.75 - 1.00	
Polar	Arg	46	54*	40	9+	149
	Lys	29+	93	119*	51	292
	His	39*	19	11	5+	74
	Gln	34	50	53	19+	156
	Asn	52	49	70	59	230
	Asp	72	70	66	52	260
	Glu	50	73	72	23+	218
Ambivalent	Ala	204*	62+	72	38+	376
	Pro	63*	38	36	22+	159
	Gly	-	-	-	-	-
	Thr	112*	73	54+	28+	267
	Ser	116*	57	74+	66+	313
Nonpolar	Cys	83*	11+	6+	0+	77
	Val	242*	56+	27+	7+	332
	Met	60*	11	6+	0+	77
	Ile	164*	33	6+	2+	205
	Leu	215*	37+	15+	6+	273
	Phe	132*	20+	5+	2+	159
	Tyr	101*	47	16+	1+	165
	Trp	43*	6	4	0+	53

* Favored (deviated more than 3 σ from average value)
+ Disfavored (deviated less than 3 σ from average value)

2.11 Possible domain interface of RNase-like domains

The models are shown to contain the residues whose accessibility values are within the ranges rarely occupied by the residues of the sample globular proteins. These residues, however, have significant information rather than being simple fault of the modeling. Since the modeled region is only a part of the 2nd largest subunit of RNA

polymerase II, there should be interface on the models which makes contact with other portions of the subunit. The corresponding region of the interface will be exposed into solvent in barnase. The difference in contact partners (from solvent to protein) should be reflected in the property of the amino acid residues which occupy the contact sites.

The sequence alignment of 5 RNase-like domains, including that of *Arabidopsis thaliana*, and 5 related RNases is shown in Figure 2-5. The residue sites, at which more than half ($\geqq$3) of the residues in the models have disfavored solvent accessibility, are indicated. The sites are largely occupied by hydrophobic amino acids and the residues are exposed at the surface of the models. Some of them correspond to the sites, at which amino acids appeared in the RNases and the RNase-like domains are largely different in their side chain properties (Fig.2-5). It suggests that the alteration in contact partners, from solvent to accompanied portions of the RNA polymerase subunit, have prompted the replacements.

Figure 2-5. Amino acid sequence alignment of the 5 RNase-like domains (*A. thaliana, S. cerevisiae, S. pombe, D. melanogaster* and *H. sapiens*) and the five RNases (Barnase, Binase, RNase T1, RNase St and RNase F1). The numbers in parenthesis indicate the residue numbers of the sequences. The results of the analysis of side chain accessibility and amino acid replacements profiles are indicated under the sequences in each row. When the types of amino acid residues which occupy a residue site are largely different between the RNase group and the RNase-like domain group, they are indicated by filled circles. The residue sites are indicated by open circles when more than half of the residues of the RNase-like domain models have disfavored accessibility.

The diagnosis of the sites is indicated at the bottom of each row. The diagnosis is categorized as follows, a: sites are mainly occupied by hydrophobic residues in the RNase-like domains but mainly hydrophilic or ambivalent (with smaller side chains) ones in the RNases, b: sites mainly by occupied by ambivalent residues in the RNase-like domains but they do not frequently appear in the RNases, c: sites replaced into hydrophobic residues in RNase-like domains, and more than half of them have disfavored accessibility.

Figure 2-6. The residue sites which have proper accessibility are viewed on the RNase-like domain model of *S. cerevisiae* in space filling model. The sites of the model are colored when more than half of the residues of 4 RNase-like domains have proper accessibility (sites which are not indicated by open circles in Fig. 2-5). The sites which are conservative among the RNases and the RNase-like domains are colored in green or magenta. When the sites are not conservative, they are colored in yellow. The putative active sites are colored by magenta. top; viewed from the active center side. bottom; opposite side.

Figure 2-7. The residue sites which have rarely observed accessibility are viewed on the model of the RNase-like domain of *S. cerevisiae* in space filling model. The sites which are indicated by '*a*', 'b' and 'c' in Fig.4, are colored in dark yellow, light yellow and green, respectively. The N- and the C-terminal residues of the model are shown in navy blue and red, respectively. top; viewed from the active center side. bottom; opposite side.

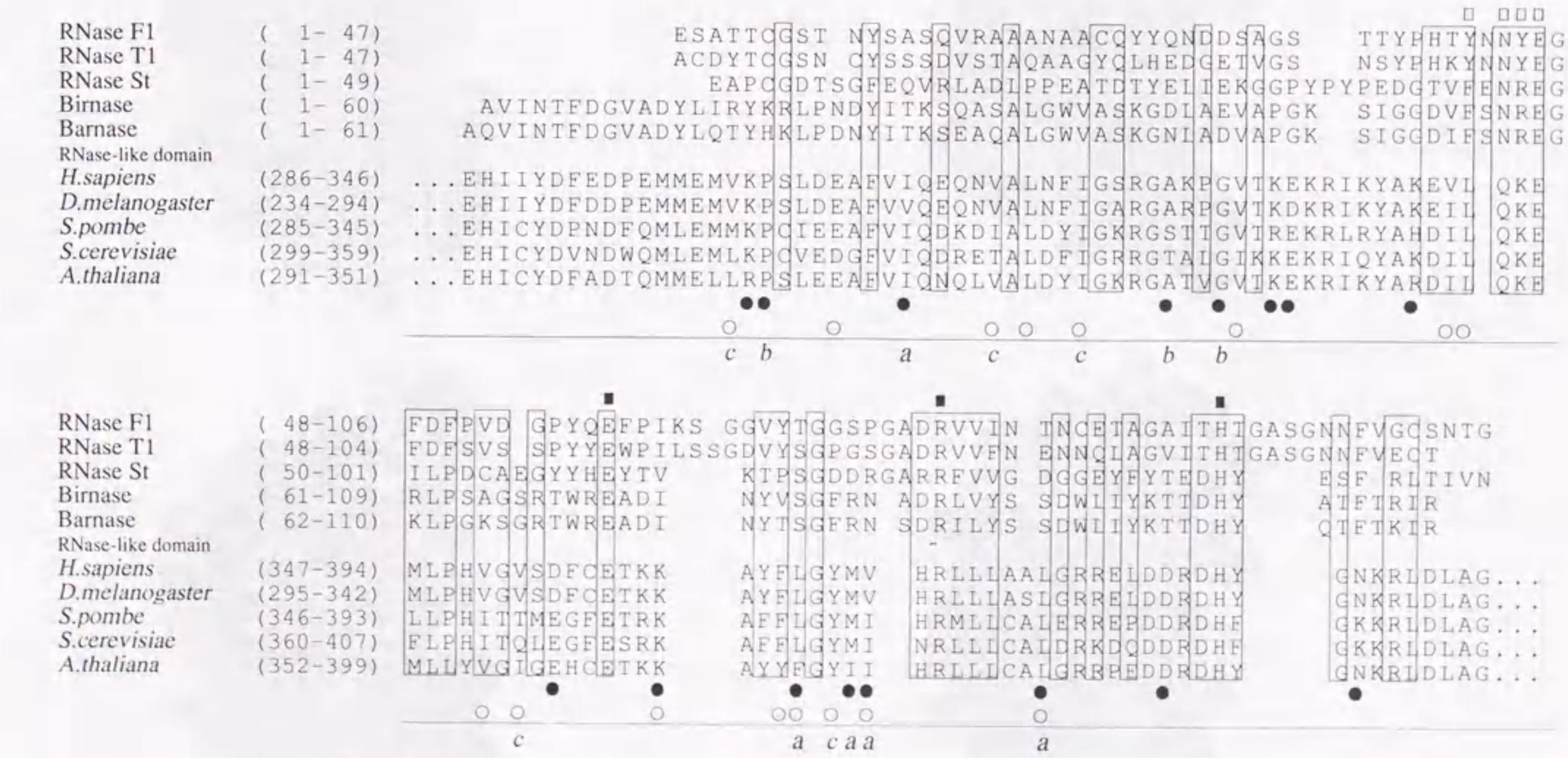


Figure 2-5 (See page 28 for legend)

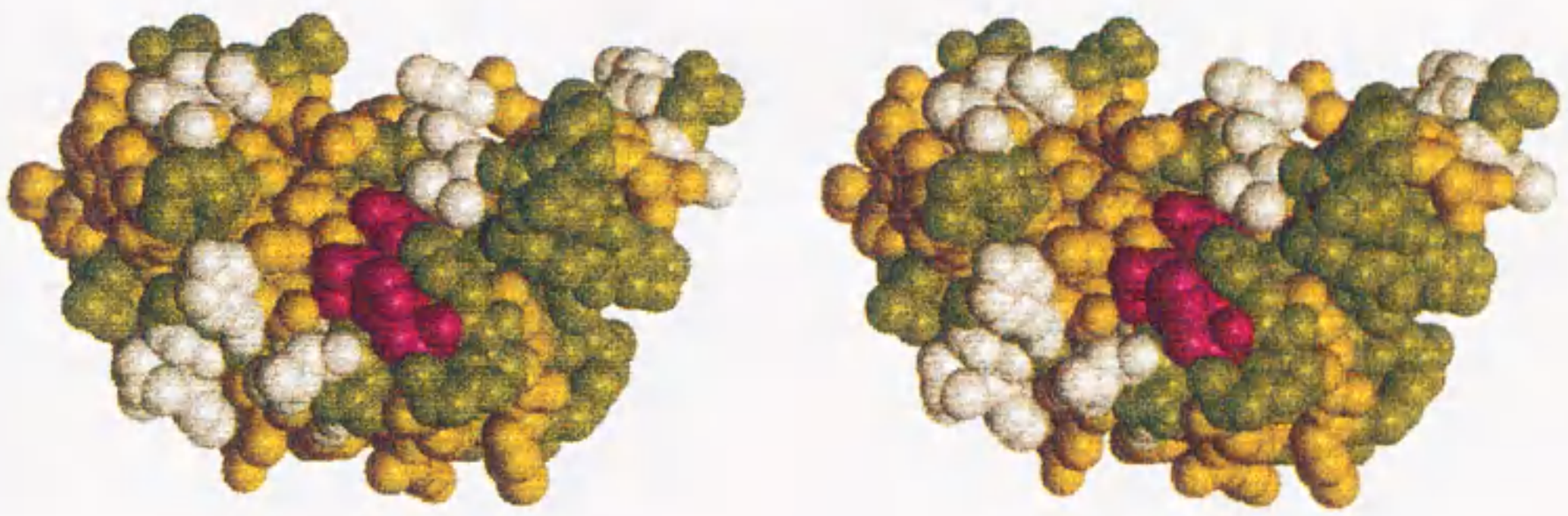


Figure 2-6 (See page 28 for legend)

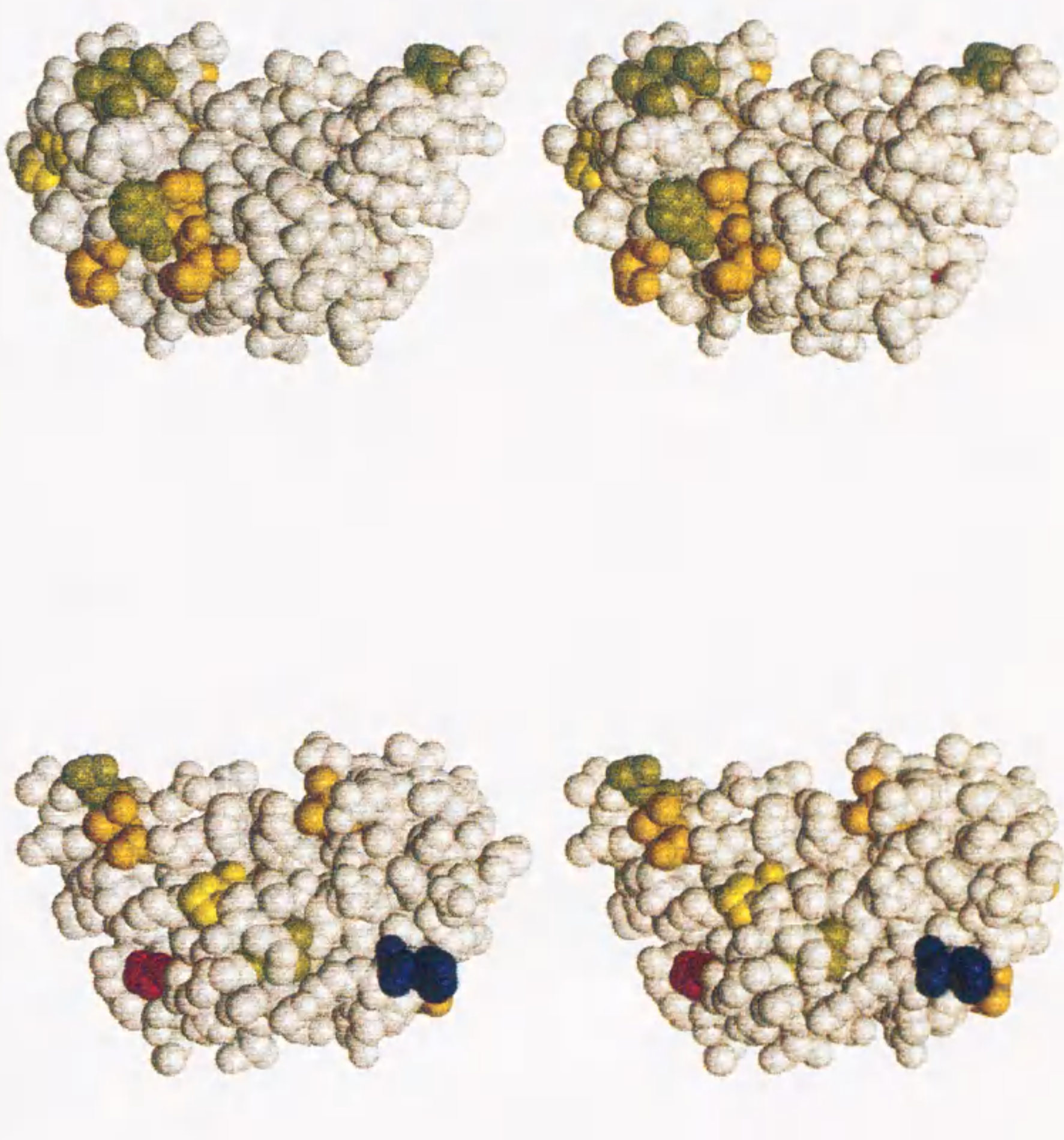


Figure 2-7 (See page 28 for legend)

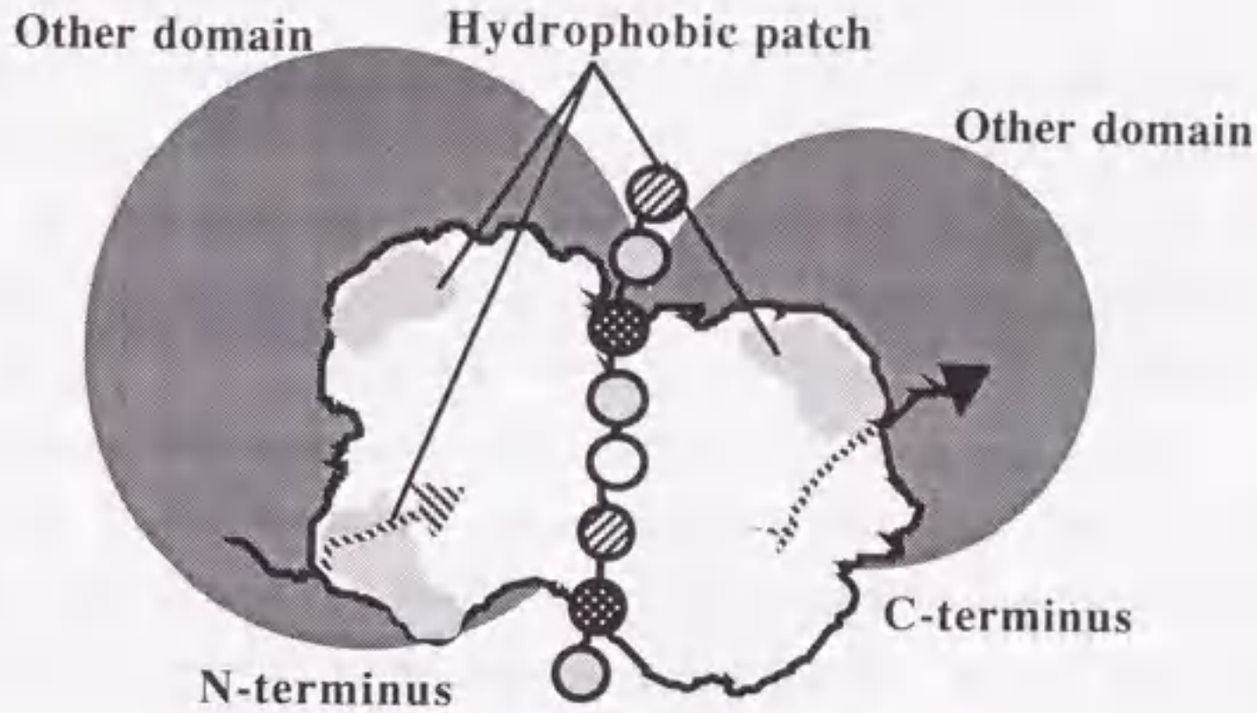


Figure 2-8 Schematic representation of the suggested inter-domain contact of the RNase-like domain. The major clusters of hydrophobic residues may be buried at interface between RNase-like domain and its N- and C-terminal domains.

It is suggested that the regions of the RNase-like domains which are in contact with other part of the RNA polymerase subunit are detected by the solvent accessibility analysis and the observation of amino acid replacements between the RNases and the RNase-like domains. The sites indicated in the alignment (Fig. 2-5) are shown on the model of *S. cerevisiae* in Figure 2-7. They are mainly clustered in two regions on the models. The regions are located near the N- or C-termini of the model. Since there are other portions of the RNA polymerase subunit in both N- and C-terminals to the RNase-like domain, the both terminals of the models are needed to contact with the other parts in native conformation. The clusters of the exposed hydrophobic residues on the surface of the model suggest that the regions take part in the inter-domain contact (Fig.2-8).

2.12 Suggested RNase Activity of the Eukaryotic RNA polymerase subunits

Recently, S-glycoproteins, which are thought to be involved in a gametophytic self-incompatibility system were found to have weak sequence similarity with fungal RNase T_2 produced by *A. oryzae* (Kawata *et al.* 1990). Although their sequence identity is only about 11%, extensive similarity around the catalytic sites of RNase T_2 suggested that S-glycoproteins are also RNases. In fact, they were found to have RNase activity (McClure

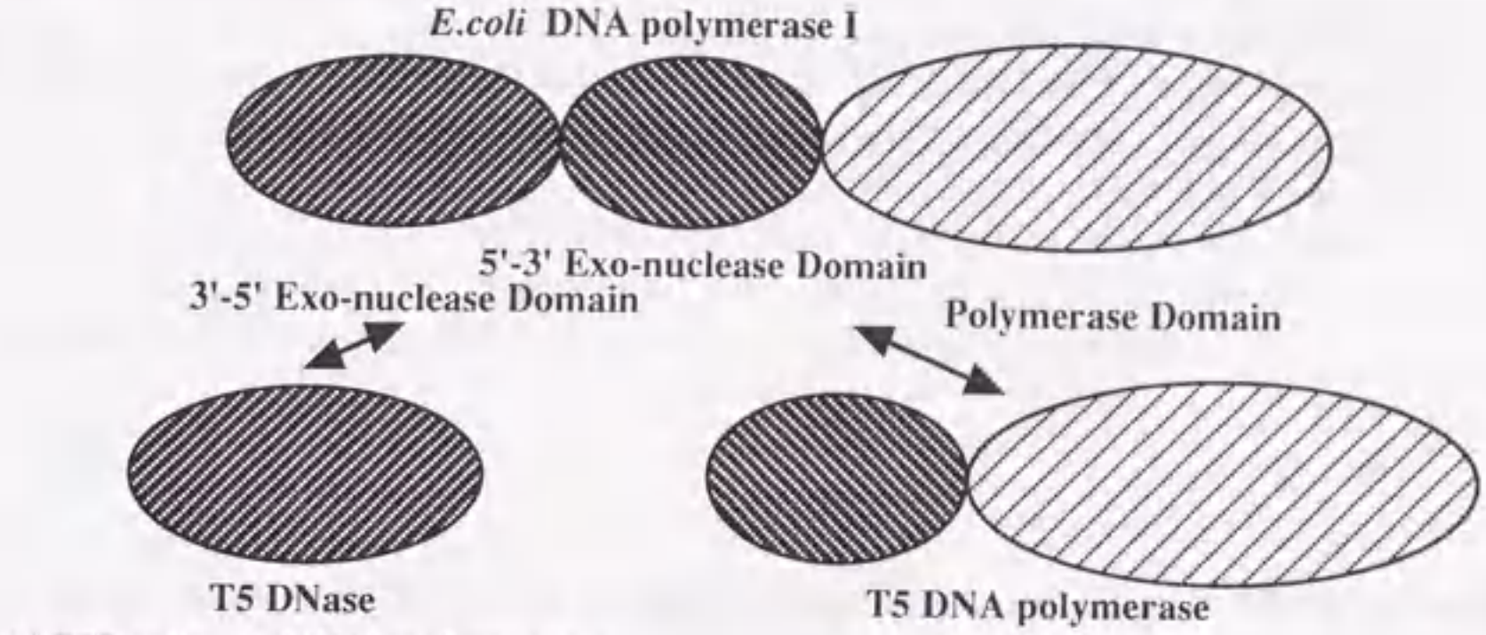
et al. 1989).

Similarly the present study suggests that the eukaryotic RNA polymerase subunits might have RNA cleavage activity. If they do, what is its contribution to the RNA catalytic process? The question should be discussed in connection with DNA polymerase, which is an essential enzyme in another nucleotide polymerizing activity. DNA polymerase I of *E. coli* has 3'→5' and 5'→3' exo-nuclease activities as well as polymerase activity (Fig.2-1). The nuclease activities have been proved to serve an essential role in DNA metabolism through increasing the fidelity of DNA replication by proofreading, facilitating the excision of defected DNA, and effecting the removal of primer fragments (Kornberg 1980). The active sites for these functions have been thought to be distributed over different regions of the enzyme. Among the activities of the enzyme, 5'→3' exo-nuclease activity can be physically isolated from the others by proteolytic cleavage. Furthermore, study of the crystal structure of the large carboxyl-terminal (Klenow) fragment of the cleaved DNA polymerase I has indicated that the active sites of the polymerase and 3'→5' exo-nuclease activities are located in two distinct domains (Ollis *et al.* 1985). These facts bear an analogy to the implications of the present study. The possession of an RNase-like domain that is distinct from a polymerase catalytic center plausibly implies that RNA polymerase also has nuclease activity since DNA polymerase possesses a similarly distinct domain (Fig.2-1). The suggested RNase activity in the RNA polymerase may serve as proofreading function in RNA polymerization activity.

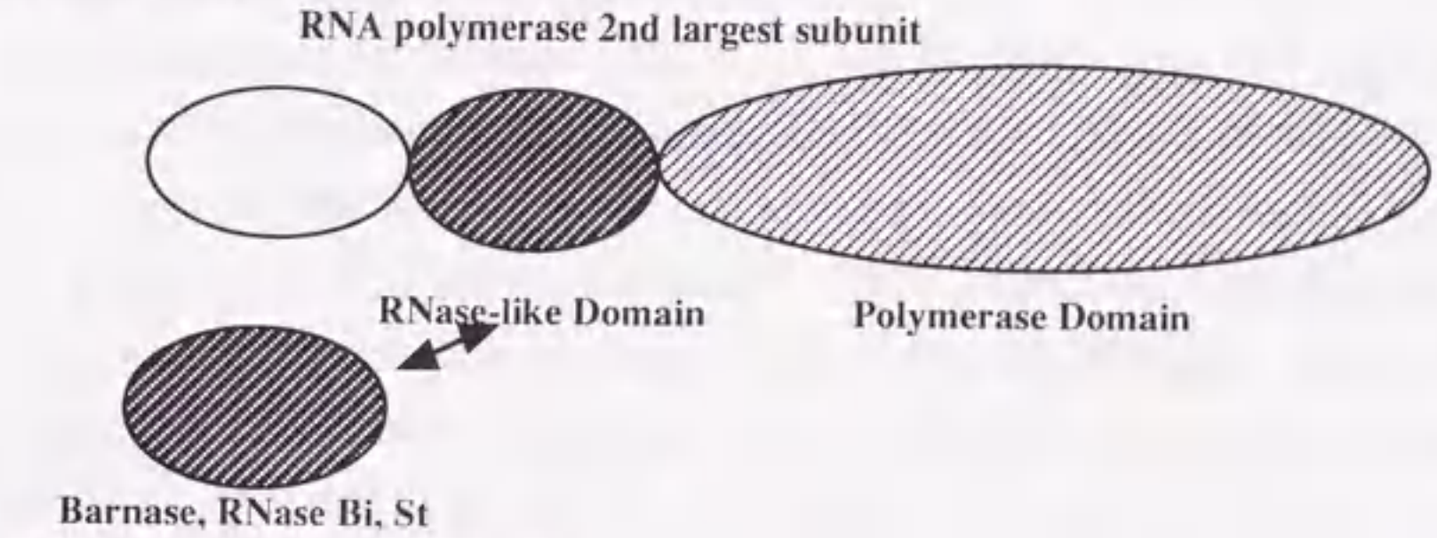
Proofreading of RNA polymerase has been found in RNA-directed RNA polymerase of influenza virus (Ishihama *et al.* 1986). Removal of excess GMP residues other than the 1st residue is carried out by the polymerase in the transcription process. The influenza virus-associated RNA polymerase also cleaves capped RNA from host cells and utilizes the resulting capped RNA fragments as primers (Plotch *et al.* 1981). This RNase activity works in an endo-nucleolytic manner, similar to that of the microbial RNases. Further experimental study should verify whether or not the RNase-like domain of the second largest subunit of RNA polymerase II has any functional role, such as RNase activity, RNA splicing and processing.

Recently, cleavage of nascent RNA molecule during transcription by RNA polymerase is reported for several organisms, such as *E. coli* (Surratt *et al.*1991, Borkhov *et al.* 1992 and 1993), vaccinia virus (Hangler and Shuman 1993), rat (Reines 1992, Reines *et al.* 1993) and human (Wang and Hawley 1993). Though the functional

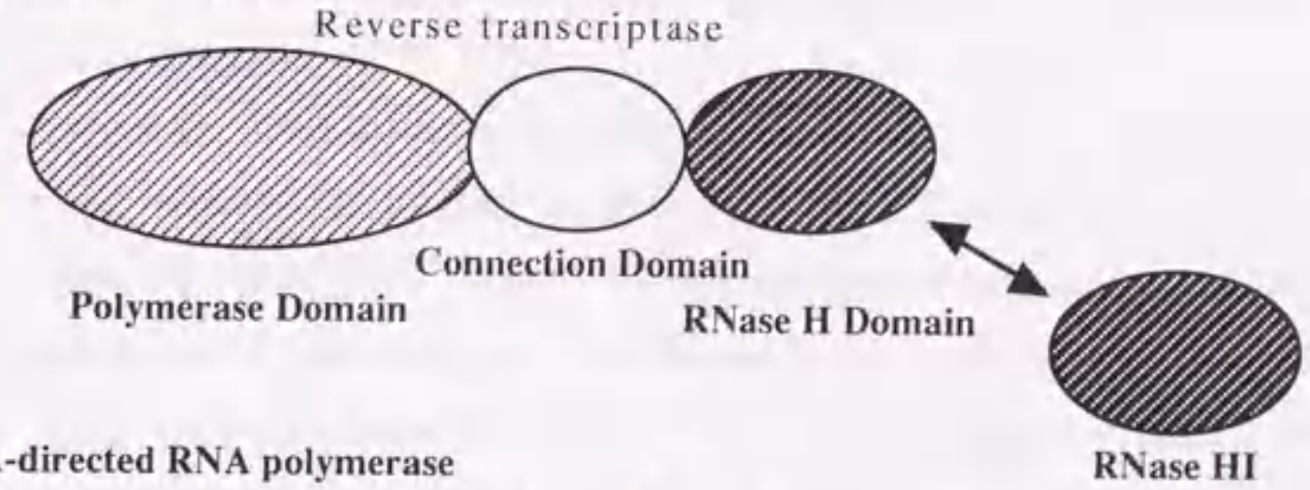
a) DNA-directed DNA polymerase



b) DNA-directed RNA polymerase



c) RNA-directed DNA polymerase



d) RNA-directed RNA polymerase

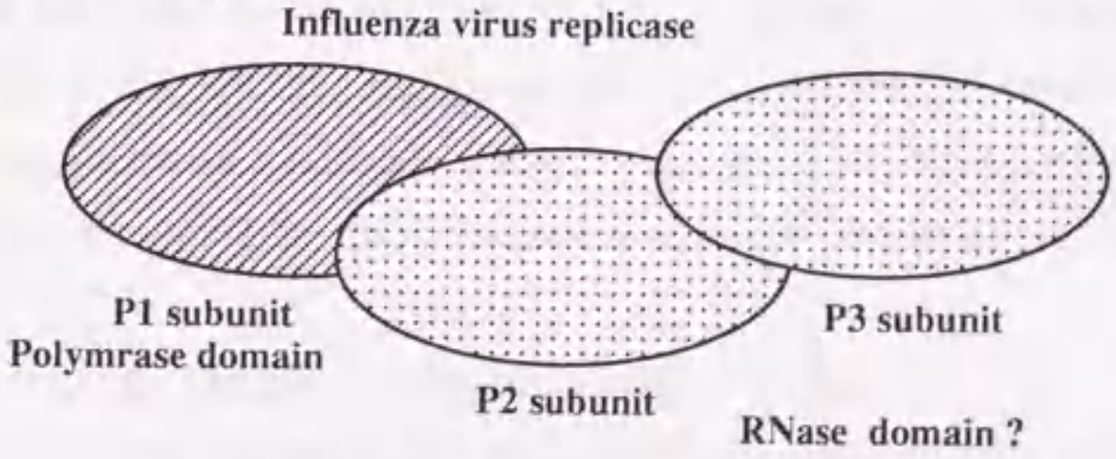


Figure 2-9. (See next page for legend)

Figure 2-9. Schematic representation of relationships between nucleotide polymerases and nucleases. **a.**DNA polymerase I of *E. coli* is composed of two C-terminal exo-nuclease domains and N-terminal polymerase domain. The homologous polymerase of T5 phage lack the counter part of the most N-terminal nuclease domain. However, T5 exo-nuclease, which is encoded by a gene different from that of the polymerase, is homologous with the domain and substitute the function of the domain. (Kornberg and Baker 1992) **b**. The RNase-like domain of RNA polymerase II 2nd largest subunit, which is found in the present study. **c.** Reverse transcriptase has a globular domain of RNase H activity. The homologous enzyme of the domain, which is named as RNase HI, have been found in various organisms. (Kohlstaedt *et al.* 1992) **d.** Nuclease activity was found in the RNA-directed RNA polymerase of influenza virus, which is composed of P1, P2 and P3 (or NP) subunits (Ishihama *et al.* 1986). The subunit(s) which bear the activity is not identified.

roles of the abortive transcription are still remain unknown, roles in recovery from dead-end ternary complex and abortive initiation are suggested.

If those functions are the main roles of the activity, they belong to a new category of nuclease activity associated with polymerases. The catalytic center of the RNase activity is still unknown. The activity is reported to require transcriptional factors SII (Reines 1992, Reines *et al.* 1993, Wang and Hawley 1993), GreA, GreB (Borkhov *et al.* 1992 and 1993) or rpo30 (Hangler and Shuman 1993) to work efficiently. However the SII itself does not have RNase activity, and RNA polymerase complex without the factor shows weak RNase activity. So the active center of the RNase activity may at least partially exists in the RNA polymerase subunits, and the RNase-like domain is a candidate for the center.

The finding of the RNase-like domain has confirmed the strong relationship between polymerases and nucleases in evolution. In fact, all kind of polymerases, DNA-directed DNA, DNA-directed RNA, RNA-directed DNA (reverse transcriptase) and RNA-directed RNA polymerases, have associated with nuclease activities (Fig.2-9). The relationship of nuclease and polymerase domains might be a typical examples of evolutionary process of functional and structural units combination in protein. The finding of RNase-like domain in RNA polymerase implies a scheme of molecular evolution governs the diversification process of nucleotide polymerases, in which a primordial polymerase has differentiated by recruiting the different nucleases as the functional domains (Fig.2-9).

2.13 Evolutionary Origin of the RNase-like Domain

The sequence similarity presented here suggests an evolutionary relationship between

the microbial RNases and the RNA polymerase subunits. One possible interpretation of this finding is that they diverged before the origin of eukaryotes. Two alternative evolutionary origins are also possible: the RNase gene was derived from a partial duplication of an ancestral RNA polymerase gene or a duplicate of the RNase gene was integrated into the latter. Also, it is conceivable that exon shuffling has brought a similar sequence into the RNase and RNA polymerase at an early stage of protein evolution. Introns have not been found at the boundaries of the RNase-like domains in yeast or fruit fly genes of the second largest subunit of the RNA polymerase; however, their ancestral gene might have had introns at module boundaries and most of these been lost, as is the case with other proteins (Go and Nosaka 1987).

A part of the sequences of the three archaebacterial B or B" subunits, two eubacterial β subunits and five higher plant chloroplast β subunits of the RNA polymerase were aligned with the RNase-like domains of the second largest subunits of the RNA polymerases of *D. melanogaster* and *S. cerevisiae* (Fig.2-2). One of the catalytic sites of barnase, His-102, is conserved in all the sequences aligned. However, Arg-87, which is the other catalytic site of barnase, is not conserved in the bacterial and chloroplast sequences, some of which were replaced with Lys or His. Glu-73 is shared in all the bacterial RNA polymerase subunits, but three of the five chloroplast subunits have Gln instead of Glu. Thus, in the prokaryotic RNA polymerase subunits not all the active sites of RNases are conserved. This fact implies that the RNase activity, which the eukaryotic RNA polymerases have been suggested to have, might have been lost in the prokaryotic RNA polymerases. However, the existence of the invariant His among all the sequences shows a possibility of another common function among them. The prokaryotic subunit sequences show weak sequence similarity with the RNases at carboxyl-terminal 59 residues of barnase, but they apparently have no residues corresponding to the amino-terminal 51 residues.

The fact that only the eukaryotic RNA polymerase subunit and not the prokaryotic one retains the catalytic sites of the microbial RNases is puzzling. As far as RNase is concerned, no obvious homolog of the microbial RNases has been found in RNases of higher eukaryotes. The combination of these facts suggests a possible complimentation of the RNase activity. The idea that the prokaryotic RNases compliment the inactivated RNase-like domain of prokaryotic RNA polymerase seems attractive, though the RNases are extra-cellular proteins and are not localized in the nucleus region. Otherwise, the loss of the catalytic sites of RNase in prokaryotic and chloroplast RNA polymerase subunits

may imply that only a part of the function of RNase (RNA binding, for example) is still incorporated in the RNA catalytic activity of the subunits.

Since the RNA world seems to have preceded the DNA world (Darnell and Doolittle 1986), RNA polymerase was presumably required prior to DNA polymerase. As the ability to pass on genetic information was essential for prebiotic evolution, RNA polymerase activity must have been among the earliest catalytic activities. It is possible that the proofreading of contemporary DNA polymerases had already existed in ancestral RNA polymerase. The presence of RNase activity in influenza virus RNA polymerases and the involvement of RNase H, as well as RNA-dependent DNA polymerase and DNA-dependent DNA polymerase, in the reverse transcriptase of retroviruses (Varmus 1982) suggest these may be remnants of such ancient RNA polymerase. However, the possibility of later acquisition of the RNase activity in the virus-type polymerases remains. In this case the coding sequences of the RNases might have been brought into the viruses as copies of their host RNases.

2.14 Bibliography

Allison, L.A., Moyle, M., Shales, M. and Ingles, C.J. 1985. *Cell* 42: 599-610.

Aphanasenko, G.A., Dudkin, S.M., Kaminir, L.B., Leshchinskaya, I.B. and Severin, E.S. 1979. *FEBS Lett.* 97: 77-80.

Barker, E.N. 1988. *J. Mol. Biol.* 203: 1071-1095.

Berghofer, B., Krockel, L., Kortner, C., Truss, M., Schallenberg, J. and Klein, A. 1988. *Nucleic Acids Res.* 16: 8113-8128.

Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C. and Kraut, J. 1982. *J. Biol. Chem.* 257: 13650-13662.

Borkhov, S., Polyakov, A., Nikiforov, V. and Goldfarb, A. 1992. *Proc. Natl. Acad. Sci. USA* 89: 8899-8902.

Borkhov, S., Savitov, V. and Goldfarb, A. 1993. *Cell* 72: 459-466.

Bott, R., Ultsch, M., Kossiakoff, A., Graycar, T., Katz, B. and Power, S. 1988. *J. Biol. Chem.* 263: 7895-7906.

Chambon, P. 1975. *Annu. Rev. Biochem.* 44: 613-638.

Chothia, C.and Lesk, A.M. 1987. *Cold Spring Harbor Symp. Quant. Biol.*, 52: 399-405.

Church, W.B., Guss, J.M., Potter, J.J. and Freeman, H. 1986. *J. Biol. Chem.* 261: 234-237.

Darnell, J.E. and Doolittle, W.F. 1986. *Proc. Natl. Acad. Sci. USA* 83: 1271-1275.

Dauter, Z. Sieker, L.C. and Wilson, K.S. 1992. *Acta. Crystallogr. sect. B* 48: 42-59.

Declercq, J.-P., Tinant, B., Parello, J. and Rambaud, J. 1991. *J. Mol. Biol.* 220: 1017-1039.

Dijikustra, B.W., Kalk, K.H., Hol, W.G.J. and Drenth, J. 1981. *J. Mol. Biol.* 147: 97

Falkenburg, D., Dworniczak, B., Faust, D.M. and Ekkehard, K.F.B. 1987. *J. Mol. Biol.* 195: 929-937.

Fujinaga, M., Delbare, L.T.J., Brayer, G.D. and James, M.N.G. 1985. *J. Mol. Biol.* 183: 479-502.

Fukuyama, K., Matubara, H. and Rogers, L.J. 1992. *J. Mol. Biol.* 225: 775-789.

Go, M. and Nosaka, M. 1987. *Cold Spring Harbor Symp. Quant. Biol.* 52: 915-924.

Grachev, M.A., Lukhtanov, E.A., Mustaev, A.A., Zaychikov, E.F., Abdukayumov, M.N., Rabinov, I.V., Richter, V.I., Skoblov, Y.S. and Chistyakov, P.G. 1989. *Eur. J. Biochem.* 180: 577-585.

Hangler, J. and Shuman, S. 1993. *J. Biol. Chem.* 268: 2166-2173.

Hartley, R.W. and Barker, E.A. 1972. *Nature (New Biol)* 235: 15-16.

Herzberg, O. 1991. *J. Mol. Biol.* 217: 701-719.

Hill, C., Dodson, G., Heinemann, U., Saenger, W., Mitsui, Y., Nakamura, K., Borisov, S., Tischenko, G., Polyakov, K. and Pavlovsky, S. 1983. *Trends Biochem. Sci.* 8: 364-369.

Hiratsuka, J., Shimada, H., Whittier, R., Ishibashi, T., Sakamoto, M., Mori, M., Kondo, C., Honji, Y., Sun, C.- R., Meng, B.-Y., Li, Y.-Q., Kanno,A., Nishizawa, Y., Hirai, A., Shinozaki, K. and Sugiura, M. 1989. *Mol. Gen. Genet.* 217: 185-194.

Hu, J. and Bogorad, L. 1990. *Proc. Natl. Acad. Sci.USA* 87: 1531-1535.

Hudson, G.S., Holton, T.A., Whitfeld, P.R. & Bottomley, W. 1988. *J. Mol. Biol.* 200: 639-654.

Ishihama, A., Mizumoto, K., Kawakami, K., Kato, A. and Honda, A. 1986. *J. Biol. Chem.* 261: 10417-10421.

Katayanagi, K., Miyazawa, M., Matsushima, M., Ishikawa, M., Kanaya, S., Nakamura, H., Ikehara, M., Matsuzaki, T. and Morikawa, K. 1992. *J. Mol. Biol.* 223: 1029-1052.

Kawata, Y., Sakiyama, F., Hayashi, F. and Kyogoku, Y. 1990. *Eur. J. Biochem.* 187: 255-262.

Kimura, M. 1983. in The Neutral Theory of Molecular Evolution. (Cambridge Univ. Press, Cambridge, England), pp. 149-193.

Kornberg, A. 1980. in DNA Replication (Freeman, San Francisco), pp. 101-166.

Leffers, H., Gropp, F., Lottspeich, F., Zillig, W. and Garrett, R.A. 1989. *J. Mol. Biol.* 206: 1-17.

Li de la Sierra, I., Papamichael, E., Sakarelos, C., Dimicoli, J.-L. and Prange, T. 1990. *J. Mol. Recog.* 3: 36

Lisitsyn, N.A., Monastyrskaya, G.S. and Sverdlov, E.D. 1988. *Eur. J. Biochem.* 1777: 363-369.

Martinez-Oyanedel, J. Choe, H.-W., Heinemann, U. and Sanger, W. 1991. *J. Mol. Biol.* 222: 335

Mauguen, Y., Hartley, R.W., Dodson, E.J., Dodson, G.G., Bricogne, G., Chothia, C. and Jack, A. 1982. *Nature (London)* 297: 162-164.

McClure, B.A., Haring, V., Ebert, P.R., Anderson, M.A., Simpson, R.J., Sakiyama, F. and Clarke, A.E. 1989. *Nature (London)* 342: 955-957.

Morikawa, K., Matsumoto, O., Tsujimoto, M., Katayanagi, K., Ariyoshi, M., Doi, T., Ikehara, M., Inaoka, T. and Ohtsuka, E. 1992. *Science* 256: 523-526.

Ohme, M., Tanaka, M., Chunwongse, J., Shinozaki, K. and Sugiura, M. 1986. *FEBS Lett.* 200: 87-90.

Ollis, D.L., Brick, P., Hamlin, R., Xuong, N.G. and Steitz, T.A. 1985. *Nature (London)* 313: 762-766.

Ovchinnikov, Y.A., Monastyrskaya, G.S., Gubanov, V.V., Guryev, S.O., Chertov, O.Y., Modyanov, N.N., Grinkevich, V.A., Makarova, I.A., Marchenko, T.V., Polovnikova, I.N., Lipkin, V.M. and Sverdlov, E.D. 1981. *Eur. J. Biochem.* 116: 621-629.

Pai, E.F., Krengel, U., Petsko, G.A., Goody, R.S., Kabsch, W. and Wittinghofer, A. 1990. *EMBO J.* 9: 2351-2359.

Plotch, S.J., Bouloy, M., Ulmanen, I. and Krug, R.M. 1981. *Cell* 23: 847-858.

Puhler, G., Lottspeich, F. and Zillig, W. 1989. *Nucleic Acids Res.* 17: 4517-4534.

Ramanadham, M., Sieker, L.C. and Jensen, L.H. 1987. *Acta. Crystallogr. sect. A* 43: 13

Reines, D. 1992. *J. Biol. Chem.* 267: 3795-3800.

Reines, D., Ghanouni, P. Li, Q. and Mote, J. Jr. 1992. *J. Biol. Chem.* 267: 15516-15522.

Riva, M., Carles, C., Sentenac, A., Grachev, M.A., Mustaev, A.A. and Zaychikov, E.F. 1990. *J. Biol. Chem.* 265: 16498-16503.

Sacchettini, J.C., Gordon, J.I. and Banaszak, L.J. 1989. *Proc. Natl. Acad. Sci. USA* 86: 7736-7740.

Smith, J.L., Corfield, P.W.R., Hendrickson, W.A. and Low, B.W. 1988. *Acta. Crystallogr. sect. A* 44: 357-368.

Stehle, T. and Schulz, G.E. 1992. *J. Mol. Biol.* 224: 1127-1141.

Surratt, C.K., Milan, S.C., Chamberlin, M.J. 1991. *Proc. Natl. Acad. Sci. USA* 88: 7983-7987.

Sweetser, D., Nonet, M. and Young, R.A. 1987. *Proc. Natl. Acad. Sci. USA* 84: 1192-1196.

Teeter, M.M. 1984. *Proc. Natl. Acad. Sci. USA* 81: 6014-6018.

Tilton Jr., R.F., Dewan, J.C. and Petsko, G.A. 1992. *Biochem.* 31: 2469-2481.

Tronrud, D.E., Holden, H.M. and Matthews, B.W. 1987. Science 235: 571-574.

Umesono, K., Inokuchi, H., Shiki, Y., Takeuchi, M., Chang, Z., Fukuzawa, H., Kohchi, T., Shirai, H., Ohyama, K. and Ozeki, H. 1988. *J. Mol. Biol.* 203: 299-331.

Van Duyne, G.D., Standaert, R.F., Karplus, P.A., Schreiber, S.L. and Clardy, J. 1991. *Science* 252: 839-842.

Van Mierlo, C.P.M., Darby, N.J., Neuhaus, D. and Creighton, T.E. 1991. *J. Mol. Biol.* 222: 353-371.

Van Roey, P.V. and Beerman, T.A. 1989. *Proc. Natl. Acad. Sci. USA* 86: 6587-6591.

Varmus, H.E. 1982. *Science* 216: 812-820.

Veerapandian, B., Gilliland, G.L., Raag, R., Svensson, A.L., Masui, Y., Hirai, Y. and Poulos, T.L. 1992. *Proteins* 12: 10-23.

Vermersch, P.S., Lemon, D.D., Tesmer, J.J.G. and Quiocho, F.A. 1991. *Biochem.* 30: 6861-6866.

Volz, K. and Matsumura, P. 1991. *J. Biol. Chem.* 266: 15511-15519.

Wang, D. and Hawley, D.K. 1993. *Proc. Natl. Acad. Sci. USA* 90: 843-847.

Wilmot, C.M. and Thornton, J.M. 1988. *J. Mol. Biol.* 203: 221-232.

Yoshida, N., Sasaki, A., Rashid, M.A. and Otsuka, H. 1976. *FEBS Lett.* 64: 122-125.

Yura, T. and Ishihama, A. 1979. *Annu. Rev. Genet.* 13: 59-97.

Chapter 3

Adaptive amino acid replacements in ribonuclease H domain accompanied by domain fusion

Protein evolution has two distinct processes, one is amino acid replacement and another is reorganization of structural or functional units of proteins. Multi-domain and multi-functional proteins are thought to have evolved by fusion of smaller functional and structural units such as modules or domains. Reverse transcriptase (RT) is one of such fused proteins. According to its 3D structure, the N-terminal part forms a globular domain which bears polymerase activity and the C-terminal part forms a globular domain with ribonuclease H activity (RNase H domain). There are homologous proteins with the RNase H domain and they exist as single domain enzymes. The group of the enzymes is called type I ribonuclease H (RNase HI). It is most likely that the ancestors of RNase HI and the polymerase domains were fused and became contemporary RT. At that time, amino acid replacements should have been occurred in fitting the fused conformation of the RNase H and the polymerase domains. Such replaced amino acid residues should have been conserved during evolution. We have compared the amino acid frequencies at each residue sites between the free form, RNase HI group, and the integrated form, RNase H domain group. It was shown that drastic fitting replacements of amino acid residues occurred at the interface of the domains; hydrophilic amino acid residues of free form were substituted with hydrophobic or ambivalent ones in the integrated form at 4 out of 29 residue sites involved in inter-domain contact. These substitutions seem to contribute to stabilize the fused conformation mainly by hydrophobic interactions at the interface of the domains. The result implies that domain fusion could have occurred by relatively small number of adaptive amino acid substitutions.

3.1 Introduction

Modular structures of proteins seem to reflect the history of protein evolution. Multi-domain proteins had been generated by fusions of smaller structural or functional units (*e.g.* Rossmann *et al.* 1975, Moras 1992, Feller 1994). Accumulation of primary and tertiary structure information of proteins has lead to a recognition that a number of proteins can be divided into smaller functional and structural units such as domains or modules (*e.g.* Rossmann *et al.* 1975, Go 1981, Holm and Sander 1994). For example, aminoacyl-tRNA synthetases are composed of several globular domains which are associated with different functions (Moras 1992). The ATP binding domains of seryl-tRNA synthetase of *E. coli* and aspartyl-tRNA synthetase of yeast are similar in 3D structures and amino acid sequences. However, these two aminoacyl-tRNA synthetases also have domains which are different each other in 3D structure and amino acid sequence. These specific domains seem to take part in recognition of their cognate tRNAs or amino acids. The 3D structural organization of these proteins imply that the biological specificity has been evolved by integrating structural units which have different functions.

Two distinctive processes, reorganization of functional/structural units and amino acid replacements, exist in protein evolution. A study of protein evolution focusing on a relation between these two process is required. Adaptive process of integrated domains into new molecular conformation would be one of the most important problems in evolutionary mechanisms of proteins. Primary and tertiary structures of RNase HI and RNase H domains of retroviral RTs were compared to study the adaptive process.

RNase HI and RNase H domain have similar biological functions. RNase HI is globular single domain ribonuclease which specifically degrades RNA strand in DNA-RNA hybrid (Stein and Haugen 1969). They are involved in DNA replication process through maturation of Okazaki fragment (Turchi *et al.* 1994). RNase H domains which correspond to the carboxyl terminal part of retroviral RTs, are homologs of RNase HI (Johnson *et al.* 1986). RNase H domain degrades RNA template during retroviral replication process (Gilboa *et al.* 1979; Goff 1990).

The crystal structure of both RNase HI (Katayanagi *et al.* 1990; Yang *et al.* 1990) and RNase H domain (Davies 1991; Kohlstaedt *et al.* 1992; Jacobo-Molina *et al.* 1993; Rodgers *et al.* 1994) were reported (Fig. 3-1).

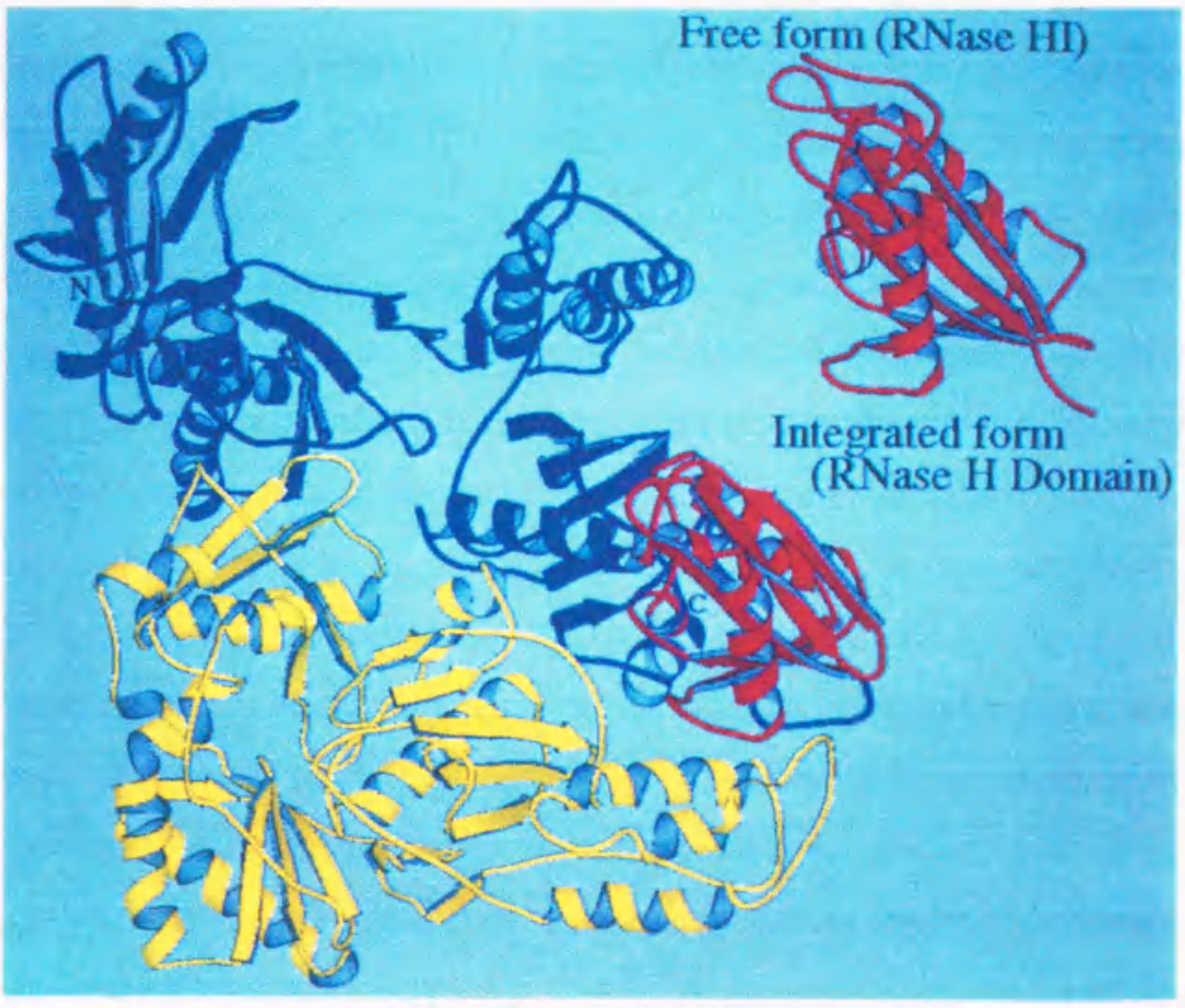


Figure 3-1 (See next page for legend)

Figure 3-1. The schematic 3D structures of reverse transcriptase and RNase HI. The p51 subunit is colored in yellow, p66 subunit except for RNase H domain is in blue. RNase H domain and RNase HI are colored in red. This figure is generated by MOLSCRIPT (Kraulis 1991).

RT is a dimer of p66 and p51 subunits. The p66 subunit is composed of three domains; polymerase domain, connection domain and RNase H domain, in the direction from N- to C-terminals. The p51 subunit which has only polymerase and connection domains, is produced by removing the RNase H domain from the product of the same gene as p66 subunit. So the p66 and p51 subunits have identical primary structure except for the absence of RNase H domain in the latter. The 3D structure of the RNase H domain is similar with that of RNase HI as expected from the amino acid sequence similarity.

Since RNase HI is a single domain enzyme and RNase H domain is the integrated form. These proteins with 3D structures known are suitable for a study of domain integration. A domain integration event would not be a simple patchwork of genes. Rather it involves accompanied fitting modifications. The fitting of a domain into integrated form would involve replacements of amino acid residues at the domain interface which is formed by the integration event. The amino acid residues at the interface are obligated to change their partners of interactions from solvent molecules to the atoms of associated domain. Replacements of amino acid residues are needed since the requirements of the physicochemical property at the interface is changed. Once the replacements have occurred, then the property of replaced residues should be conserved. The conservation of the residues during the course of evolution would be detected by the residues conserved among the diverged offsprings. The residues of the free form which correspond to the conserved residues at the interface of integrated form are more free for replacements. The amino acid fractions at each corresponding residue site between the groups of five free forms, RNase HI group, and eight integrated forms, RNase H domain group were compared. The amino acid frequencies of residue sites reflect the substitution patterns of the sites. By the comparison of the frequencies, the patterns of amino acid replacements and the atomic interactions of the replaced residues were studied.

Figure 3-2. a Amino acid sequence alignment of RNase HIs and RNase H domains of RTs. The top five sequences are RNase HIs (free form), and the others are RNase H domains of RTs (integrated form) in each row. The source of the sequences is in the text. Numbers in parenthesis indicates residue numbers. For the RNase H domains, residue numbers are counted from the amino terminal of polyprotein, except for HIV1 which is counted from the proteolytic cleavage site to keep a consistency with the numbering system for the crystal structure of HIV1 RT (Jacobo-Molina *et al.* 1993). The amino terminal of RNase HI of *S. cerevisiae* is not completed and the number was counted from known amino terminal. The conserved sites (d(i)=0) are boxed in thick line. The sites accounted in amino acid composition analysis are boxed in a thin line. The sites appeared in Table III-I are indicated by filled circle (type A), filled triangle (type B), hatched square (type C) and filled square (type D). The basic protrusion is boxed by dotted line. The residual sites take part in inter-domain contact in RNase H domain are indicated under each row by (*) and (@), which stand for those from accounted sites and non-accounted sites, respectively. **b** A stereo drawing of back bone traces of the superimposed structures of RNase H domain and RNase HI. The traces are presented in tube models. The RNase H domain is colored in yellow and the RNase HI is in blue. A part of the structure of RNase HI which is extended upward is the basic protrusion. **c** Distances between superimposed C$^{\alpha}$s of RNase-H domain of HIV1 RT and RNase HI of *E. coli*. The open circles are the distances between all aligned sites (see **a** of this figure) of the two sequences. The residue numbers are presented in the numbering system of RNase HI. The total superimposed residues are 122 sites and the RMS distance is 3.6 Å. The sites under the line for 3.0Å are selected as accounted sites. The filled circles are the distances obtained by a superimposition of the accounted sites. The number of the accounted sites are 106, and the RMS distance is 1.6Å.

It was shown that most drastic change in side chain property occurred at the interface of domains. Twenty-nine residue sites of RNase H domain are involved in inter-domain contact. They were identified by the 3D structures of RNase H domain with its associated domain and RNase HI using solvent accessibility and atomic contact at the domain interface. At least 4 sites out of the 29 inter-domain contact sites were drastically changed their side chain property from hydrophilic to hydrophobic or ambivalent. It seems they contribute in stabilizing the fused conformation mainly by hydrophobic interactions.

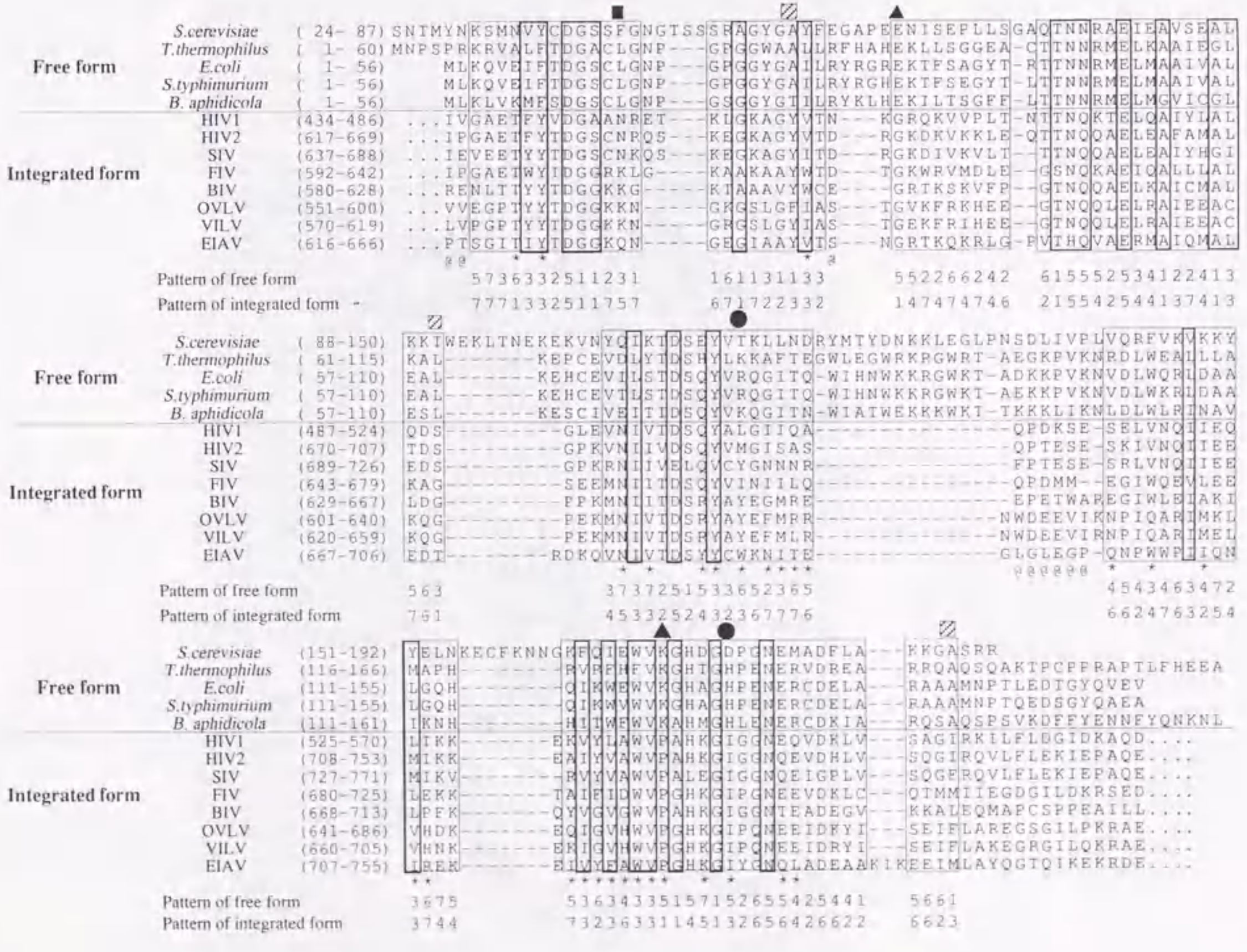


Figure 3-2a (See page 45 for legend)

Figure 3-2b (See page 45 for legend)

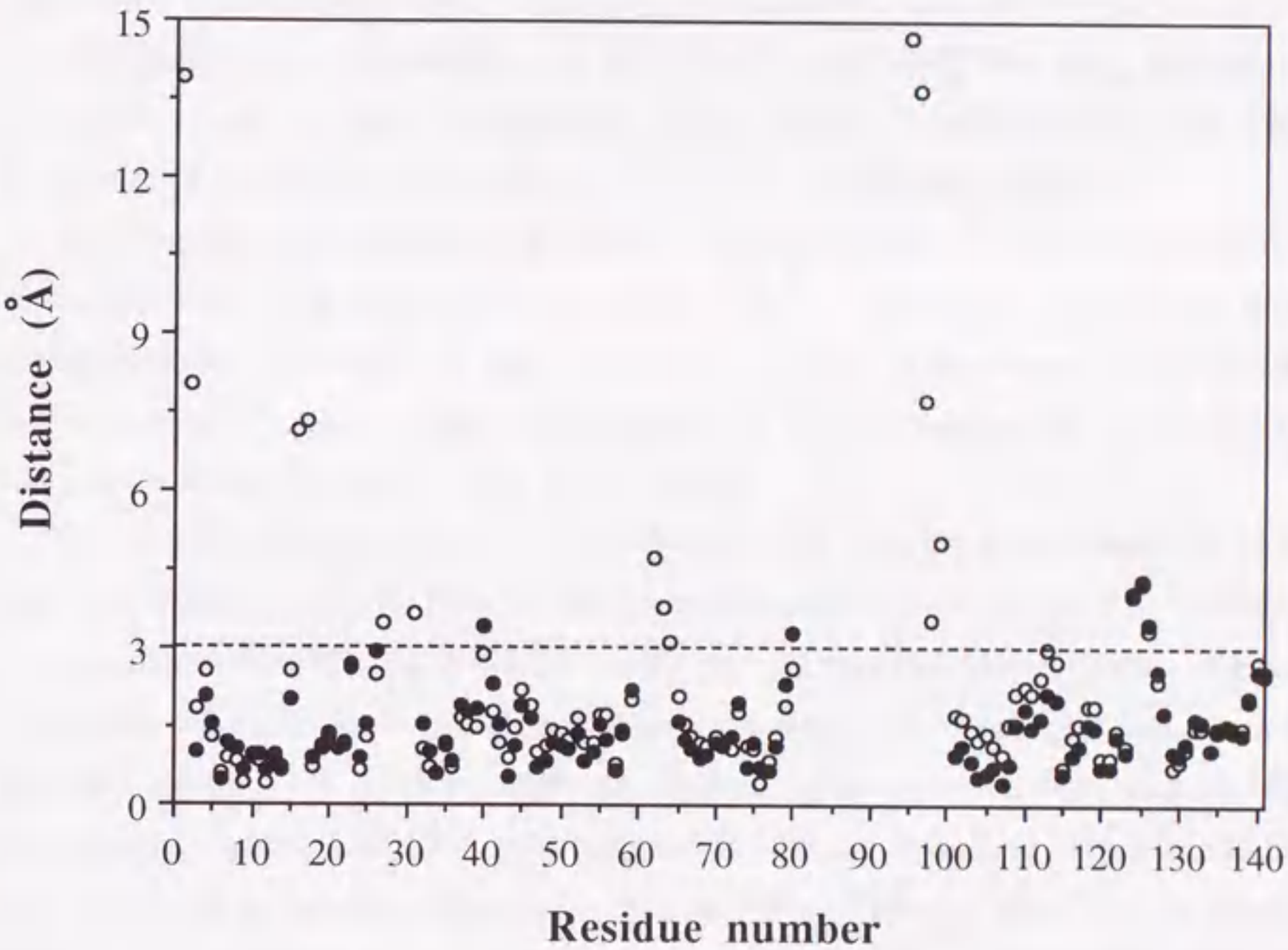


Figure 3-2c (See page 45 for legend)

3.2 Alignment of RNase H domains

Thirteen amino acid sequences of RNase H domains were obtained from SwissProt data base (release 30.0). The sources of sequences are as follows: RNase HI sequences are from *Esherichia coli* (Kanaya and Crouch 1983), *Salmonella typhimurium* (Itaya *et al.* 1991), *Thermus thermophilus* (Itaya and Kondo 1991), *Buchnera aphidicola* (Munson *et al.* 1993) and *Saccharomyces cerevisiae* (Itaya *et al.* 1991). The retroviral RNase H domains of RT are from human immunodeficiency virus type1 (HIV1) (Ratner *et al.* 1985) and type2 (HIV2) (Guyader *et al.* 1987), simian immunodeficiency virus (SIV) (Franchini *et al.* 1987), feline immunodeficiency virus (FIV) (Talbott *et al.* 1989), bovine immunodeficiency virus (BIV) (Garvey *et al.* 1990), ovine lentivirus (OVLV) (Querat *et al.* 1990), visna lentivirus (VILV) (Sonigo

et al. 1985) and equine infectious anemia virus (EIAV) (Stephens *et al.* 1986). The coordinates of the crystal structures were obtained from Brookhaven National Laboratory Protein Data Bank (Bernstein *et al.* 1977). The 3D structures used are RNase HI of *E. coli* (Katayanagi *et al.* 1992); BNL code 2RN2, the integrated form of RNase H domain of HIV1 (Kohlstaedt *et al.* 1992; Jacobo-Molina *et al.* 1993; Rodgers *et al.* 1995); BNL codes of them were 1HVT, 1HMI and 1HMV.

The alignment was obtained by IDEAS (Kanehisa 1982) (Fig. 3-2). Since the roles of residue sites were analyzed on the basis of the 3D structures, equivalence of the spatial positions of the sites of free and integrated forms of the domain was needed to be confirmed. It was done by superimposition of 3D structures of the free form of *E. coli* and the integrated form of HIV type1 (HIV1).

From the 122 superimposed sites, 103 were selected for the further analysis by the condition that the distances between the superimposed C^{α}s were within 3Å. The three continuous residues, His124 to Gly126 of RNase HI and His539 to Gly541 of RNase H domain, were also included, though they don't satisfy the condition. The reason is that they are the only residues which are flanked by the selected sites without being interrupted by gaps. The 106 selected residue sites are boxed in Fig.3-2, and they will be referred as accounted sites in the present study. The root mean square distance between superimposed C^{α}s was 1.6Å for the 106 accounted sites. The active site residues of RNase H, Asp10, Glu48, Asn130, His124 and Asp134, in the numbering system of *E. coli* RNase HI, were included in the accounted sites (Jacobo-Molina and Arnold 1992).

3.3 Detection of adaptive replacements accompanied by integration

To detect adaptive replacements, amino acids were clustered into three classes (Go and Miyazawa 1980): hydrophilic (Asp, Asn, Glu, Gln, Arg, Lys and His), hydrophobic (Val, Leu, Ile, Cys, Met, Phe, Tyr and Trp) and ambivalent ones (Gly, Ala, Ser, Thr and Pro). The pattern of appearance of the classes (Fig. 3-3) at each residue sites was compared between the two forms of the domain. A quantitative criterion was introduced for pattern comparison. Based on the alignment (Fig. 3-2), fractions of the classes at each residue site within the free forms, RNase HIs (5 sequences), were calculated and the same within the integrated forms (8 sequences) were also calculated, after which the fractions were compared. Differences in the fractions between the free and integrated forms at the ith residue site were detected as

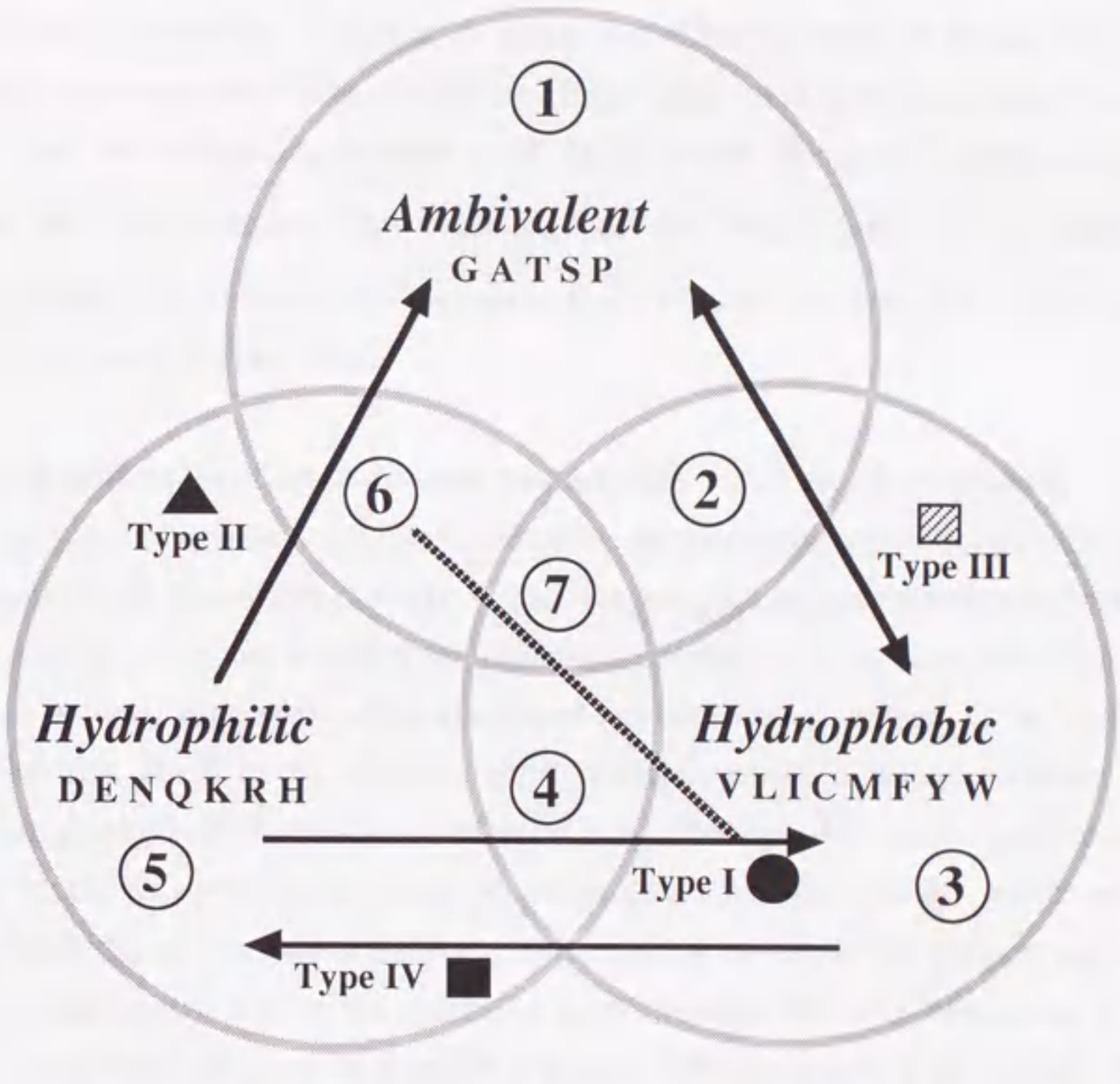


Figure 3-3. Schematic representation of amino acid appearance and replacement patterns. Amino acids within a circle belong to the same class. The patterns of amino acid class appearance are designated by numbers. Three classes and their combinations make 7 patterns. The arrows and the associated symbols indicate replacement patterns from free to integrated domain forms, or between.

$$d(i)=1/2\sum_{k}^{3}(a_{ik}-b_{ik})^2, \qquad (1)$$

where a_{ik} and b_{ik} are the fractions of class k (k=1, 2, 3) at the ith residue site of the free form and integrated form, respectively. The value of $d(i)$ is between 0 and 1. If

*i*th residue site has the same compositions between the two forms, then $d(i)$ is 0. When the class is conserved within each group but different between groups, then $d(i)$ is 1. Adaptive replacement sites should have high values of $d(i)$, because adaptively replaced residues are expected to be conserved. So the value of b_{ik} of a certain class is higher than for other classes. However, a_{ik} of the same class is not high, because replacements in the free form are easier. Eight residue sites with a $d(i)$ value higher than 0.84 are listed in Table III-I.

3.4 Identification of inter-domain contact sites based on 3D structures

Inter-domain contact sites were identified by solvent accessibility of side chains and distance from the partner domain. A site was assigned as inter-domain contact site when the residue at the site satisfied the following conditions. 1) at least one atom of its side chain is water accessible when associated domains are removed. 2) at least one atom of the side chain exists within a short distance, where a water molecule cannot be accommodated in, from atom(s) of associated domains. Accessible surface area of the side chain was calculated using the documented method (Shrake and Rupley 1973). Accessibility of the side chain was obtained by dividing the surface area in native conformation by that in its extended conformation (Go and Miyazawa 1980). The solvent accessibility can be compared among different amino acid residues, because it is normalized by the maximum accessible surface area and it takes on a value between zero and one.

The number of adaptively replaced sites was discussed together the number of inter-domain contact sites. The number of the contact sites of RNase H domain was compared with those obtained from domains of other proteins. Four multi-domain proteins which have domains comparable with size of the RNase H domain, were used. An averaged fraction of residue sites buried at the domain interface was compared with that of the RNase H domain. Crystal structures of N-terminal domains of papain (Drenth *et al.* 1976), thermolysin (Holmes and Matthews 1982) and rhodanese (Ploegman *et al.* 1978) and VL domain of Fab fragment of immunoglobulin (Lascombe *et al.* 1989) were used to obtain numbers of inter-domain contact sites by applying the same criterion as used for RNase H domains.

3.5 Pattern of amino acid class appearance in free and integrated domains

The appearance pattern of the amino acid classes at residue sites was compared

between free and integrated domain forms (Fig.3-2). There were several sites at which the classes are largely altered between the two forms. The d(*i*) scores for each site were calculated for a quantitative comparison.

The top 8 sites shown in Table III-I, according to *d*(*i*) values, have drastic replacement patterns. The amino acid classes of these sites have been completely conserved in integrated forms and the same class is never observed in the free forms (Table III-I).

Table III-I Candidates of adaptive replacements for domain fusion in reverse transcriptase

Residue No.		*d*(*i*)	Pattern[a]		Type[b]	Contact partners in p51 subunit	Adaptive replacement
H-domain	HI		H-domain	HI			
Asn447	Leu14	1.00	5	3	IV		
Tyr457	Ala24	1.00	3	1	III		
Gly462	Glu32	1.00	1	5	II		yes
Ser489	Leu59	1.00	1	3	III		
Pro537	Lys122	1.00	1	5	II	Gly262, Asn265, Val261	yes
Ile542	His127	1.00	3	5	I	Gln258, Val261, Cys280, Leu283	yes
Ile556	Ala141	1.00	3	1	III		
Leu503	Arg75	0.84	3	6	I	Pro421	yes
Thr459	Leu26	0.77	2	3	*no*		
Val552	Ala137	0.77	2	1	*no*		
Lys454	Gly21	0.67	7	1	*no*		
Thr439	Glu6	0.64	1	6	*no*	Ala288, Leu289	
Glu438	Val5	0.58	7	3	*no*		

[a] See Fig. 3-3 for definition.

[b] See text or Fig. 3-3 for definition.

Preference for amino acid classes at certain sites changed by domain integration. The 8 sites were classified into 4 types, as follows (Fig. 3-3): when hydrophilic residues of the free forms were replaced with hydrophobic or ambivalent residues in the integrated forms, the sites belong to type I or type II, respectively. Sites at which replacement between hydrophobic and ambivalent residues occurred, belonged to type III. Sites at which hydrophobic residues were in the free forms but hydrophilic ones in the integrated forms, belong to type IV (Table III-I). To determine whether the observed class switch was accompanied by domain integration events, it was checked to see if the sites are involved in the domain interface.

3.6 Adaptive replacements at domain interface

Type I (Arg75 and His127 in *E.coli* RNase HI) and type II (Glu32 and Lys122) sites located at the domain interface, when the free form structure is superimposed on the integrated from (Fig. 3-4). Domain interface sites are close to the associated domain in the integrated form but exposed on the surface in the free form. The proximity of a site to the associated domain was measured by the distance between its C^{α} atom and the nearest C^{α} atoms of the associated domain. Exposure of a site to surface in a free form was represented by solvent accessibility of the side chain at the site (Fig. 3-5). The interface sites were within a short distance from the associated domain but did have a large accessibility in a single form.

Figure 3-4. **a.** Identified adaptive evolution sites viewed on the structure of *E. coli* RNase HI superimposed onto RNase H domain of RT. The space filling model is *E. coli* RNase HI (Katayanagi *et al.* 1992) which is superimposed to RNase H domain of HIV1 RT (Jacobo-molina *et al.* 1993). A Portion of RT, except for RNase H domain, is presented in the tube model. The presented regions are from Asp237 to Pro433 of p66 subunit (yellow), and from Ile244 to Tyr427 of p51 subunit (green). The side chains of the adaptive replaced sites (Table III-I) are colored; type A in red, type B in magenta and type D in green. Type C sites are hardly visible from outside. **b.** The same structure as **a**. is rotated to the left about 90° around up-down ward axis on paper. The region in which the adaptive replaced sites concentrate is circled. **c.** The adaptive sites in the RNase H domain and the contact partners of the associated domains. The portion of RT crystal structure which corresponds to the circled region in **b**. is closed up. The side chains and C^{α}s of the adaptive sites and those of the contact partners are presented in space filling model. The main chain trace around the sites are presented in the tube model. The coloring is coordinated with the **a** and **b**..

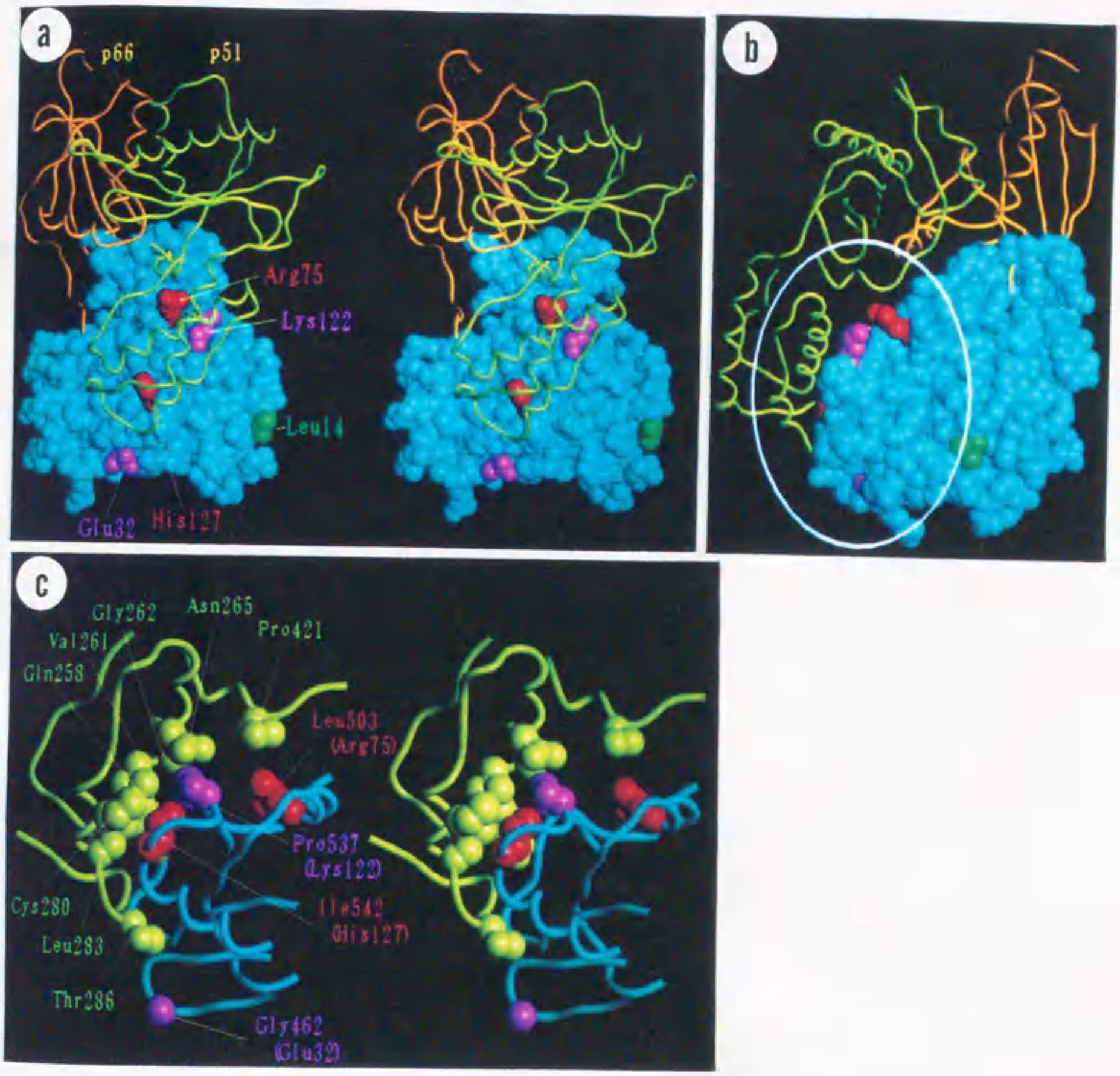
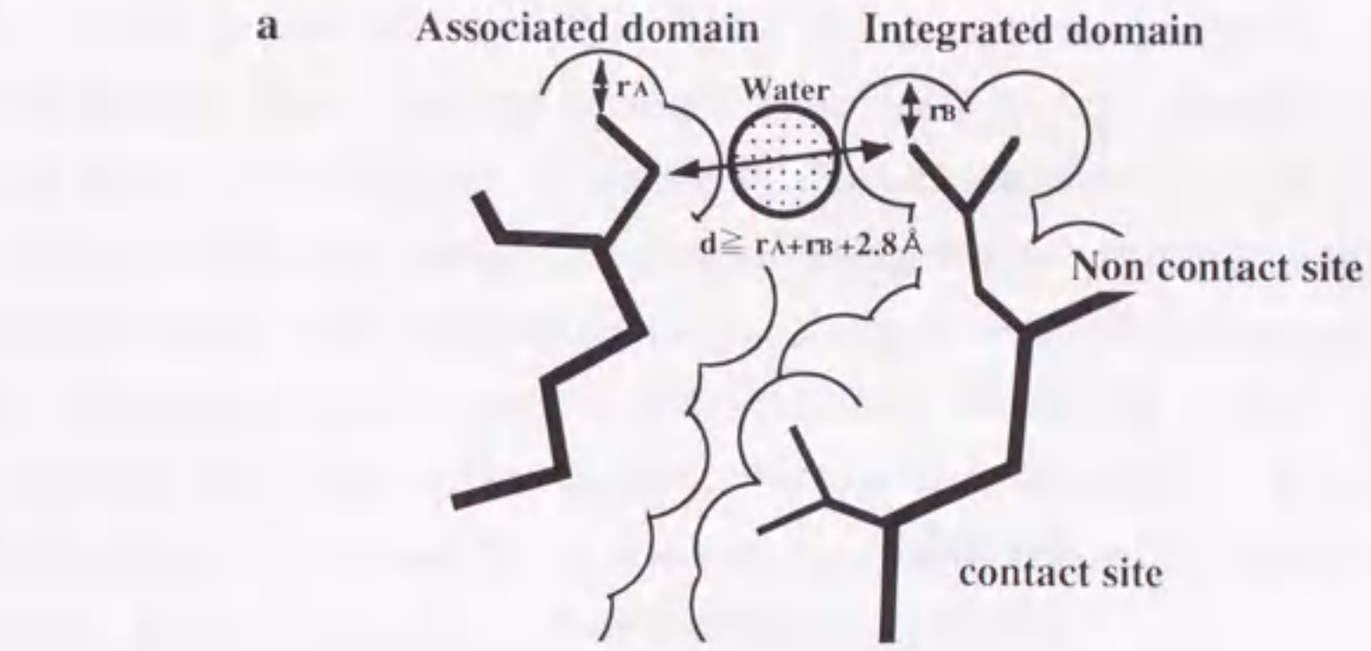


Figure 3-4 (See previous page for legend)



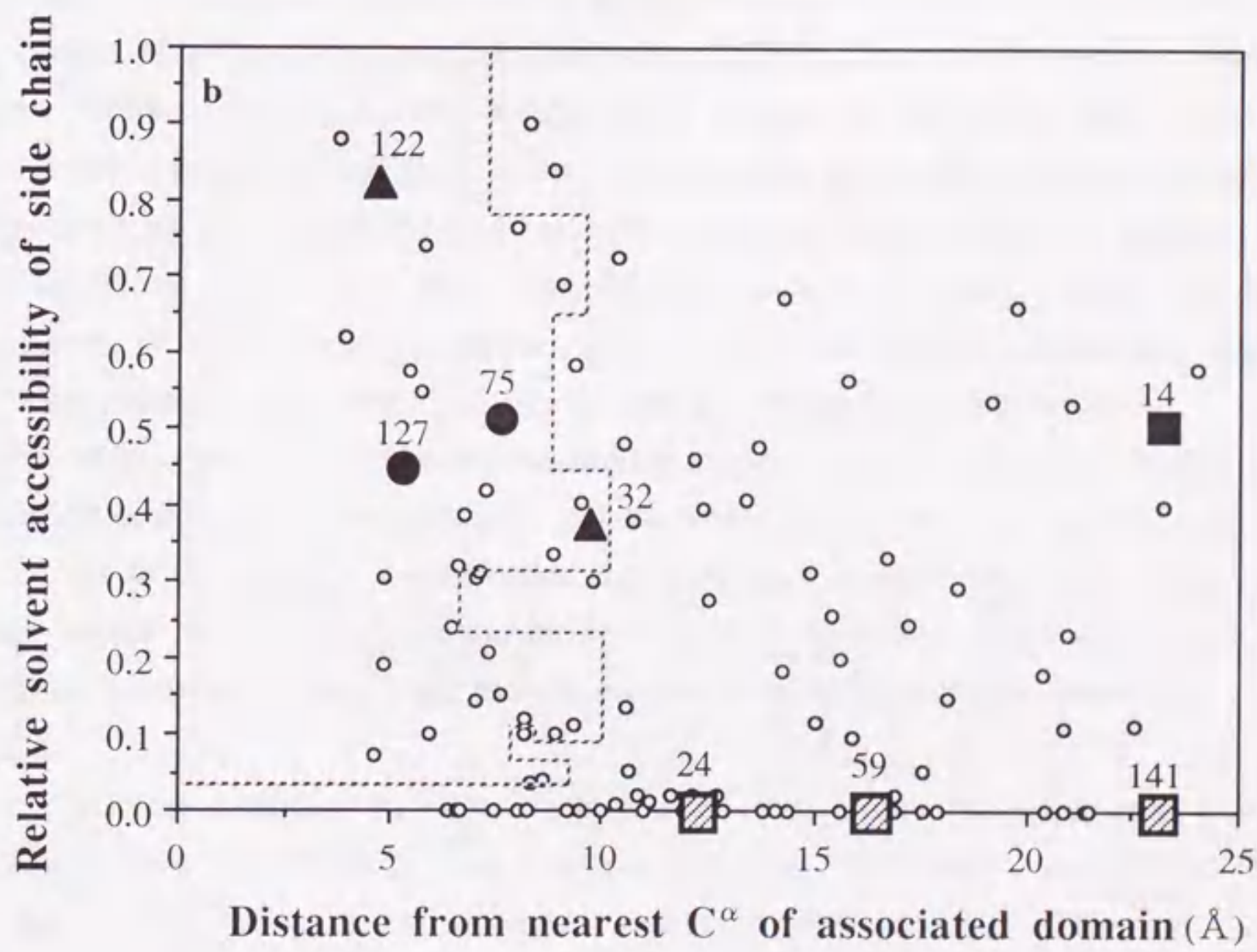


Figure 3-5. **a.** The schematic representation of the criteria of the contact sites. **b.** The relationship between side chain accessibility of the residues of RNase HI and distance from associated domain of RT when superimposed to RNase H domain. Accessibility of side chains are calculated on *E. coli* RNase HI structure. The distance from RT is between Cα of the RNase HI (superimposed onto RNase H domain) and the nearest Cα of RT other than RNase H domain. The sites appeared on Table III-I are indicated by same symbols as in Fig. 3-2 and Fig. 3-3, with residue numbers in the numbering system of *E. coli* RNase HI. The other residues are indicated by smaller open circles. The sites on the left of dotted line make contact or crash with associated domain when superimposed.

At domain integration, the most remarkable replacement would be type I or type II, as observed at sites 32, 75, 122 and 127, and exclusively occurring in regions with small distance values and large accessibility values (Fig. 3-5). Regions subjected to the most drastic change during the course of domain integration are the interface of the domain. The interface between the domains is exposed to solvents in a free form, but is buried in contact with the associated domains as a result of domain integration. Thus, type I replacement sites make for an hydrophobic interaction at the domain interface. Contact partners of these sites on associated domains are listed in Table III-I. Ile542 of RNase H domain makes for an hydrophobic interaction with Val261, Cys280 and Leu283. Leu503 makes hydrophobic contact with Pro421.

Type II replacement would obviate possible steric hindrance which might be caused by contact of the domains, or avoid burial of hydrophilic residues. The nearest contact partner of Pro537 is Gly262. Pro537 also makes hydrophobic interaction with Val261. The larger hydrophobic side chain at this site may cause steric hindrance with Gly262. The other type II site, Gly462 with no side chain, does not make contact with associated domain atoms. However, when the structure of the free form, RNase HI, is superimposed on the RNase H domain, the corresponding Glu32 makes contact with associated domain atoms (Fig. 3-5). Gly462 may possibly relax steric hindrance.

Two other types of replacement seem less significant. Type III replacement pattern is abundant within protein interiors (Go and Miyazawa 1980), and probably does not contribute to the interface, because side chains of these residues have little access to solvents and no direct effect on the surface (Fig. 3-5). Type IV replacement is distant from the associated domains and there is no direct interaction with the domains (Figs. 3-4, 5).

There is a deletion of 14 residues, called as basic protrusion (Katayanagi *et al.* 1992), in RNase HI (Fig.3-2). The deletion may also contribute in stabilization of domain location. When *E. coli* RNase HI is superimposed on RNase H domain, the protrusion overlap a part of polymerase domain (Davies II *et al.* 1991; Kohlstaedt *et al.* 1992). The protrusion of RNase HI is shown to take part in substrate binding (Kanaya *et al.* 1991). The binding function have been taken over by polymerase domain in RT (Jacobo-Molina *et al.* 1993). This takeover is thought to tolerate the loss of basic protrusion. It implies there are two distinct process in adaptation accompanied by domain fusion. One is the amino acid replacements discussed above, and the another is degeneration of dispensable functional unit. Though this might not be a necessary

process in this case because the basic protrusion structure is conserved, for example, in RNase H domain of molony leukemia virus RT (Davies II et *al.* 1991).

3.7 Fraction of adaptively replaced sites at domain interface

What is the required extent of residue replacements to adapt an integrated domain form? The ratio of the number of the adaptively replaced sites and the number of inter-domain contact sites was next calculated.

Twenty-nine residue sites of 104 accounted ones take part in the inter-domain contact in the integrated form, RNase H domain (Table III-II). This number is similar to that calculated from RNase HI superimposed on the RNase H domain (Fig. 3-4). The number of contact sites is about 27% of the accounted sites and the ratio is similar to those for four other proteins; papain, thermolysin, rhodanese and immunoglobulin. These domains are comparable in size to RNase H domain. The ratio of inter-domain contact sites in total accounted sites is 25±6% for these domains. The ratio of RNase HI or RNase H domain is within the range.

The 4 sites identified as adaptive replacement at domain integration (types I and II) are about 14% of the 29 total contact sites. The number of sites, where amino acid properties were changed remarkably to accommodate the domain integration, is not large, which means that domain integration may have occurred frequently during the evolution of proteins. The other 25 sites at the interface also take part in inter-domain interactions. However, the interactions are made by the residues which can also appear in free forms with the same properties. They were involved in the interactions without adaptive replacements.

In this study, 104 residue sites are selected out of total 138 sites of RNase H domain of HIV1 as the accounted sites. The number of total contact sites in the integrated form, RNase H domain, is 38 when the contact criteria used for the accounted sites are applied to all sites of the integrated forms. The 9 non-accounted contact sites correspond to the sites around indels between the free and the integrated forms of the domain (Fig. 3-2). They seem to be localized because the indels largely disturb 3D structure. It is possible that these structurally diverged sites take part in the adaptive evolution. However, the amino acid residues which occupy the non-accounted contact sites are not conserved within the integrated forms. Conservation of amino acid property is a part of the criteria for adaptive replacements. Since the 9 non-accounted contact sites of the integrated forms do not satisfy this criteria, the number of the

identified adaptive sites won't change, even if the structurally diverged sites are taken into account.

Table III-II Number of inter-domain contact residues in multi domain proteins

Name of proteins	Domain size (No. of residues accounted)	No. of contact residues	Fraction in accounted sites
RNase HI[a]	155 (106)	30	0.28
RNase H domain	138 (104)	29	0.28
Papain	112 (112)	37	0.33
Thermolysin	143 (143)	30	0.21
Rhodanese	156 (156)	32	0.21
Ig-VL domain	110 (110)	25	0.23
			0.25 ± 0.06[b]

[a] Superimposed on RNase H domain.

[b] Average and standard deviation of the four sample domains.

3.8 Roles of adaptive evolution at domain interface

Residue replacements at the domain interface would stabilize the multi-domain structure by increasing atomic interactions. This process involves formation of more hydrophobic interactions and increment in compatibility of domain surfaces.

The importance of the process is supported by experiments which clarified the necessity of proper positioning of catalytic sites in polymerase and ribonuclease activities and their coupled action (Oyama *et al.* 1989; Hostomsky *et al.* 1991; Post *et al.* 1993). A chimeric RT with the polymerase domain from murine leukemia virus and the artificially combined *E. coli* RNase HI produces shorter RNA cleavage products; it disrupts the primer extension probably by cutting the template RNA before it is copied, while the polymerase activity is comparable with wild type (Post *et al.* 1993). Part of the disruption can be explained by the instability of fused conformation of polymerase and RNase H domains. The domains connected by a covalent bond are apparently not sufficient for an effective functional coupling. It is likely that the

adaptive residue replacement was needed to increase atomic interactions required for domain positioning (Fig. 3-6). Other investigators reported that the RNase H domain of HIV1 is inactive when separated from the polymerase domain but regains activity when co-existing with an associated domain (Hostomsky *et al.* 1991). This interdependence of the domains can be explained by the adaptation of the RNase H domain into fused conformation.

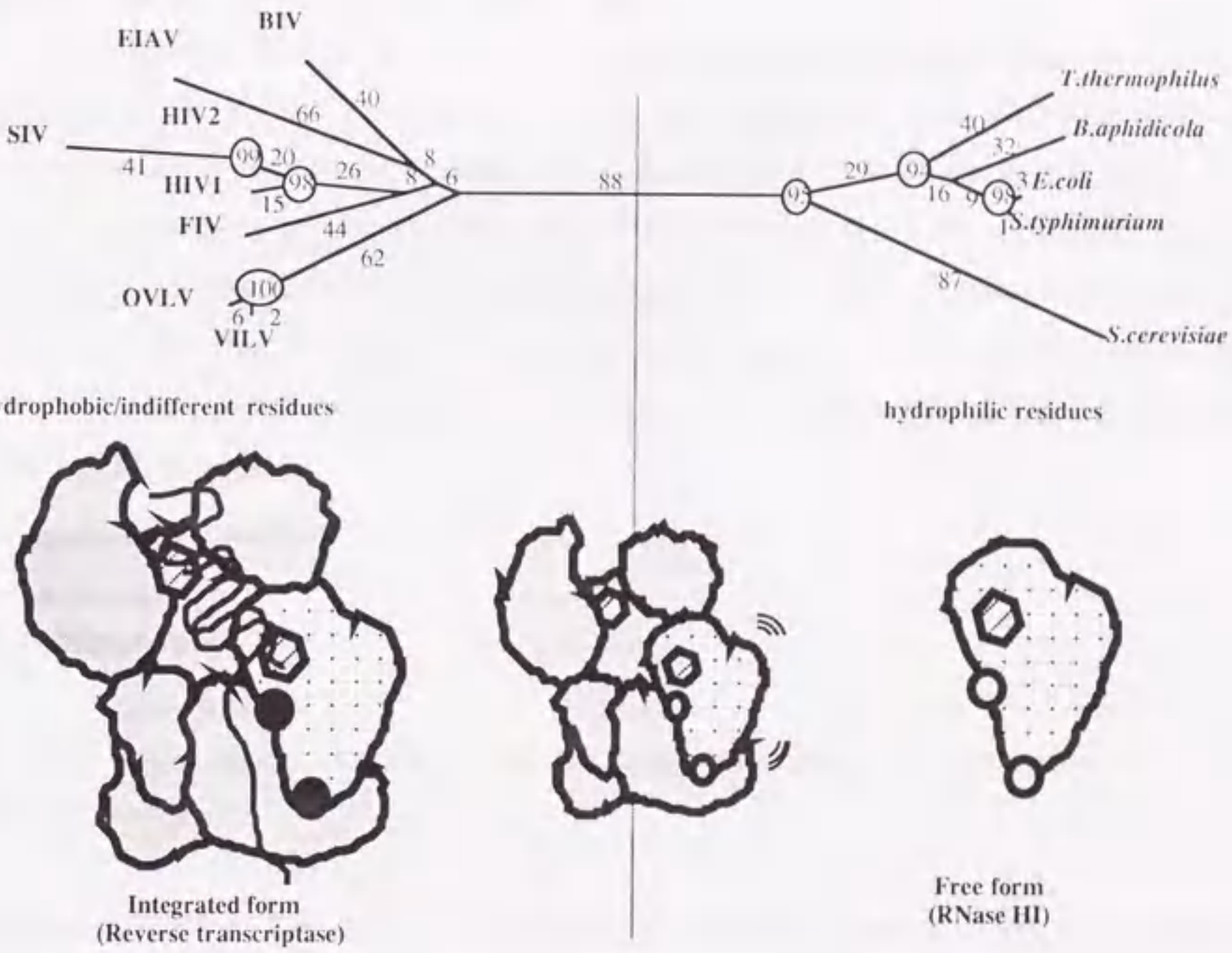


Figure 3-6. The scheme of adaptive evolution at domain interface associated with the phylogenetic relationship between RNase HIs and RNase H domains. The phylogenetic tree was constructed by neighbor joining method, using evolutional distances deduced from the alignments. The evolutionary distances were calculated by the maximum likelihood method using PROTML program (Adachi and Hasegawa 1992). The distances are presented on each branch in unit of substitution/site. The bootstrap probabilities higher than 90% are indicated on corresponding nodes. Hydrophilic residues (open circle in cartoon under the tree) of free form enzyme are replaced into hydrophobic or ambivalent residues (filled circle) at domain interface of integrated form domain. The replacements contribute the stabilization of the fused confomation and the spacing between catalytic centers (hexes in figure).

It implies that effects of amino acid replacements at domain interface can influence, through the stabilization effect in inter-domain interaction, the efficiency of the enzyme activity. Many activities of proteins are based on protein-protein interactions. The inter-domain interaction is a kind of protein-protein interaction. The result suggests that essential part of newly formed interface between two domains or proteins can be adapted by a few amino acid replacements. The most important implication of the result would be that a new complex of proteins can be obtained without much difficulty during the course of protein evolution.

It implies that effects of amino acid replacements at domain interface can be connected to the higher level character such as replication efficiency of retroviruses. Improvements in phenotypic features of organisms have their basis on improvements in molecular functions. However, it is still difficult to relate these two processes directly . Protein-protein interactions mediate the transmission of protein functions to higher level phenotypic features. It seems that studies on evolution of inter-protein interactions, including those of inter-domain interactions, contribute to morphological or functional evolution of organisms.

3.9 Bibliography

Adachi, J. and M. Hasegawa. 1992. Computer Science Monographs, no. 27. Institute of Statistical Mathematics, Tokyo.

Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Mayer, Jr, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. *J. Mol. Biol.* 112: 535-542.

Davies II, J. F., Z. Hostomska, Z. Hostomsky, S. R. Jordan, and D. A. Matthews. 1991. *Science* 252: 88-95.

Drenth, J., K. H. Kalk, and H. M. Swen. 1976. *Biochem.* 15: 3731-3738.

Feller, S. M., R, Ren, H. Hanafusa, and D. Baltimore. 1994. *Trends Biochem.* 19: 453-459.

Franchini, G., C. Gurgo, H. -G. Guo, R. C. Gallo, E. Collalti, K. A. Fargnoli, L. F. Hall, F. Wong-Staal, and M. S. Reitz Jr. 1987. *Nature* 328: 539-543.

Garvey, K. J., M. S. Oberste, J. E. Elser, M. J. Braun, and M. A. Gonda. 1990. *Virology* 175: 391-409.

Gilboa, E., S. W. Mitra, S. Goff, and D. Baltimore. 1979. *Cell* 18: 93-100.

Go, M., and S. Miyazawa. 1980. *Int. J. Peptide Protein Res.* 15: 211- 224.

Go, M. 1981. *Nature* 291: 90-92.

Goff, S. P. 1990. *J. Acquired Immune Defic. Syndr.* 3: 817- 831.

Guyader, M., M. Emerman, P. Sonigo, F. Clavel, L. Montagnier, and M. Alizon. 1987. *Nature* 326: 662-669.

Holm, L., and C. Sander. 1994. *Proteins* 19: 256-268.

Holmes, M. A., and B. W. Matthews. 1982. *J. Mol. Biol.* 160: 623-639.

Hostomsky, Z., Z. Hostomska, G. O. Hudson, E. W. Moomaw, and B. R. Nodes, 1991. *Proc. Natl. Acad. Sci. USA* 88:1148-1152.

Itaya, M., and K. Kondo. 1991. *Nucleic Acids Res.* 19:4443-4449.

Itaya, M., D. McKelvin, S. K. Chatterjie, and R. J. Crouch. 1991. *Mol. Gen. Genet.* 227: 438-445.

Jacobo-Molina, A., and E. Arnold. 1991. *Biochem.* 30:6351-6361.

Jacobo-Molina, A., J. Ding, R. G. Nanni, A. D. Clark Jr., X. Lu, C. Tantillo, R. L. Williams, G. Kamer, A. L. Ferris, P. Clark, A. Hizi, S. H. Hughes, and E. Arnold. 1993. *Proc. Natl. Acad. Sci. USA* 90:6320-6324.

Johnson, M. S., M. A. McClure, D. -F. Feng, J. Gray, and R. F. Doolittle. 1986. *Proc. Natl. Acad. Sci. USA* 83: 7648-7652.

Kanaya, S., and R. J. Crouch. 1983. *J. Biol. Chem.* 258: 1276-1281.

Kanaya, S., C. Katsuda-Nakai, and M. Ikehara. 1991. *J. Biol. Chem.* 266:11621-11627.

Katayanagi, K., M. Miyagawa, M. Matsushima, M. Ishikawa, S. Kanaya, M. Ikehara, T. Matsuzaki, and K. Morikawa. 1990. *Nature* 347: 306-309.

Katayanagi, K., M. Miyagawa, M. Matsushima, M. Ishikawa, S. Kanaya, H. Nakamura, M. Ikehara, T. Matsuzaki, and K. Morikawa. 1992. *J. Mol. Biol.* 223:1029-1052.

Kraulis, P. J. 1991. *J. Appl. Cryst.* 24: 946-950.

Kohlstaedt, L. A., J. Wang, J. M. Friedman, P. A. Rice, and T. A. Steitz. 1992. *Science* 256:1783-1790.

Lascombe, M.-B., P. M. Alzari, G. Boulot, P. Saludjian, P. Tougard, C. Berek, S. Haba, E. M. Rosen, A. Nisonoff and R. J. Poljak. 1989. *Proc. Natl. Acad. Sci. USA* 86: 607-611.

Moras, D. 1992. *Trends Biochem. Sci.* 17: 159-164.

Munson, M. A., L. Baumann, and P. Baumann. 1993. *Gene* 137: 171-178.

Oyama, F., R. Kikuchi, R. J. Crouch, and T. Uchida. 1989. *J. Biol. Chem.* 264: 18808-18817.

Ploegman, J. H., G. Drent, K. H. Kalk, and W. G. J., Hol. 1978. *J. Mol. Biol.* 123: 557-594.

Post, K., J. Guo, E. Kalman, T. Uchida, R. J. Crouch, and J. G. Levin. 1993. *Biochem.* 32: 5508- 5517.

Querat, G., G. Audoly, P. Sonigo, and R. Vigne. 1990. *Virology* 175: 434-447.

Ratner, L., W. Haseltine, R. Patarca, K. J. Livak, B. Starcich, S. F. Josephs, E. R. Doran, J. A. Rafalski, E. A. Whitehorn, K. Baumeister, L. Ivanoff, S. R. Petteway Jr., M. L. Pearson, J. A. Lautenberger, T. S. Papas, J. Ghrayeb, N. T. Chang, R. C. Gallo, and F. Wong-Staal. 1985. *Nature* 313: 277-284.

Rodgers, D. W., S. J. Gamblin, B. A. Harris, S. Ray, J. S. Clup, B. Hellmig, D. J. Woolf, C. Debouck, and S. C. Harrison. 1995. *Proc. Natl. Acad. Sci. USA*. 92: 1222-1226.

Rossmann, M. G., A. Liljas, C. -I. Brändén, and L. J. Banaszak. 1975. Pp.62-102 in P. D. Boyer eds. The Enzymes, Vol. XI. Academic Press, New York.

Shrake, A., and Ruply, J. A. 1973. *J. Mol. Biol.* 79: 351-371.

Sonigo, P., M. Alizon, K. Staskus, D. Klatzmann, S. Cole, O. Danos, E. Retzel, P. Tiollais, A. Haase, and S. Wain-Hobson. 1985. *Cell* 42: 369-382.

Stein, H., and Hausen, P. 1969. *Science* 166: 393-395.

Stephens, R. M., J. W. Casey, and N. R. Rice. 1986. *Science* 231: 589-594.

Talbott, R. L., E. E. Sparger, K. M. Lovelace, W. M. Fitch, N. C. Pedersen, P. A. Luciw, and J. H. Elder. 1989. *Proc. Natl. Acad. Sci. USA* 86:5743-5747.

Turchi, J. J., L. Huang, R. S. Murante, Y. Kim, and A. Bambara. 1994. *Proc. Natl. Acad. Sci. USA* 91: 9803-9807.

Yang, W., W. A. Hendrickson, R. J. Crouch, and Y. Satow. 1990. *Science* 249: 1398-1405.

Conclusion

The finding of the RNase-like domain in DNA-directed RNA polymerase suggests that polymerase-nuclease association is a general scheme in molecular evolution of polymerases (chapter 2). From the structures of the polymerases, they seem to be specialized by acquiring the various nucleases as functional domains in the course of evolution. The activities of the nuclease domains are used in different ways such as proofreading in DNA-directed DNA and RNA-directed RNA polymerases, degradation of RNA template in reverse transcriptase and degradation of primer in DNA polymerases. The RNA cleavage activity in DNA-directed RNA polymerase serves for a new function, abortion of transcription.

The nucleotide polymerases might be typical examples of evolutionary process of proteins by domain fusion; a single functional domain (putative primordial polymerase domain) has differentiated into complex enzyme systems by recruiting similar but different functional units (nuclease domains) for different activities. The finding of the RNase-like domain, as a prediction of functional domain structure in RNA polymerase, brings the DNA-directed RNA polymerase under the scheme.

The nuclease domains in DNA polymerase I (exo-DNase domain) and reverse transcriptase (RNase H domain) are globular domains composed of continuous polypeptide. Also the amino acid sequences of the RNase-like domains imposed on barnase 3D structure imply that they form globular domains. Continuous polypeptide and globular conformation are favorable features for a building unit of protein. The former is required because they can be inserted or deleted in one event. The latter is required to minimize the perturbation in overall conformation which is caused by insertion or deletion of a domain.

In the inspection of the amino acid sequences of RNase-like domains, several residue sites were found to have strange spatial distributions and solvent accessibility. These sites are occupied by hydrophobic/ambivalent amino acids and exposed to solvent at the surface of the domain. An assumption that the exposed residues might form the domain interface bring a question on adaptive evolution accompanied by domain fusion.

A globular domain introduced into a protein is going to be adapted to the new molecular environment. The polymerase and the nuclease activities are required to work in concert to properly synthesize nucleic acid polymer. Some adaptive amino acid replacements might be required to achieve the concerted action.

The comparison of amino acid sequences and structures of the RNase H domain of reverse transcriptase (integrated form of domain) and RNase HI (free form of domain) suggests 4 residue sites as adaptive sites (chapter 3). The residue sites have been substituted from hydrophilic amino acids in the free form into hydrophobic or ambivalent amino acids in integrated form. The residues contribute to stabilize the integrated conformation by hydrophobic interactions and removing steric hindrance at the interface. The roles of the residues are adequate ones and the substitution pattern will be generally observed in other multi-domain proteins.

The amount of the adaptive replacements is an important parameter in protein evolution. If an adaptation process requires too many replacements, it limits the opportunity for novel combinations of domains. In the case of RT, the 4 sites are only 14% of a total of 29 sites at the interface. It implies that essential part of adaptation can be achieved by replacements of few residues. It largely reduces a cost of domain fusion and may explain why the strategy of nuclease domain recruitment is widely accepted by the nucleotide polymerases.

It is important to see whether the suggestions obtained in this study are generally observable in other cases. Also there are other problems in molecular evolution by reorganization of building units of proteins. For example, frequency of such rearrangements, size distribution of building units and relationship between structural and functional diversification processes would be ones of the most important problems which are remained to be solved.

Acknowledgements

I wish to thank to Prof. Mitiko Gō for her kind and intensive instructions in this study. Also I wish to thank to Dr. Tosiyuki Noguti for his kind advises. I am grateful for Dr. Kaoru-Fukami Kobayashi, Mr. Kei Yura and all other colleagues for their helps in this study. I wish to thank to Dr. Hori for his critical reading of the manuscript. Author was supported by the Fellowship of the Japan Society for the Promotion of Science for Japanese Junior Scientists in 1993.

副論文

1. RNase-like domain in DNA-directed RNA polymerase II.
   Tsuyoshi Shirai and Mitiko Gō
   Proc. Natl. Acad. Sci. USA, 88, 9056-9060 (1991)
   （RNAポリメラーゼIIのリボヌクレアーゼ様ドメイン）

2. Adaptive amino acid replacements accompanied by domain fusion in reverse transcriptase.
   Tsuyoshi Shirai and Mitiko Gō
   J. Mol. Evol. (in press)
   （逆転写酵素におけるドメイン融合進化にともなう適応的アミノ酸置換）